Investigating the presence and impact of competing events on prognostic model research

Lucy Teece

A thesis presented for the degree of

Doctor of Philosophy

Keele University

December 2019

Abstract

Prognostic models are used to predict an individual's future health outcomes, including the risk of disease progression and the development of further complications. The statistical methodology used to develop these models is often naïve to the presence of competing events, these are events which prevent or alter the probability of an outcome of interest from occurring. Not appropriately accounting for competing events is known to produce inflated absolute risk predictions for time-to-event outcomes, this bias is known as competing risks bias. However, there has been relatively little research about competing events in prognostic model research, for which absolute risk predictions are a key outcome.

This thesis investigates the presence and impact of competing events on prognostic model research. To begin, two reviews were conducted to determine the presence, reporting, and management of competing events in current prediction model literature. Then competing risks statistical regression methods were applied to develop and internally validate a prognostic model using existing study data. These models were compared to models developed using standard time-to-event analysis techniques, naïve to competing events, with an external validation study. Finally, a simulation study was performed to identify the circumstances for which competing risks bias affects the predictive ability and calibration of prognostic models, with an overall aim to provide guidance for the optimal approaches to incorporate competing risks in prognostic model research.

Acknowledgements

This PhD would not have been possible without funding from the National Institute for Health Research (NIHR), School for Primary Care Research (SPCR), for which I am truly grateful.

I would like thank my many supervisors: **Prof. Richard Riley**; for your guidance, for embracing my independent nature, for encouraging me to pursue all of my extracurricular activities, and for teaching me how to be enthusiastic about my research without looking like a total dork. **Dr Kym Snell**; for joining the team and for making time for me when needed, and for your encouragement throughout. **Prof. Danielle van der Windt**; for your support, insight, and great input throughout my PhD. And **Dr Sally Wilkes**; for discussing and helping to shape the project in its early stages. I enjoyed working with all of you, and hope to continue to collaborate with you in the future.

I would also like to thank my colleagues and collaborators, who worked with me throughout the project. In particular **Prof. Shakila Thangaratinam**, **Nadine Marlin** and the other members of the **PREP collaboration** for agreeing to share their data for this project. I would also like to thank the patients who contributed their data to the project.

I would like to thank my fellow **PhD students**/coffee machine co-owners, who made every day a pleasure and made me want to travel to Keele every day. To **Dr Miriam Hattle**; my stats-twin and partner in crime. Your kindness and support has helped to keep me sane though the many highs and lows of the past few years. I'm so proud of what you have achieved, and hope you now see how amazing and strong you truly are. To Aunty **Lorraine (Dr Watson)**; without your dark and twisted humour the student room would have been just plain dull. Thank you for letting me bounce ideas off you, and for sharing your wisdom with me.

Finally, I would like to add a special thank you to my family and friends.

To **Mum and Mick**, for always encouraging me and making sure I believed in myself. I wouldn't have made it this far without you. To my **nieces and nephew**, for keeping me young, reminding me how to have fun, and for showing me that it is possible to have a productive day after only a few hours of sleep. To **Nana and Grandad**, my biggest fans, not a phone call has gone by without the word proud being said a trillion times. I love you all.

To my amazing friends: The **Newcastle ladies**; for distracting me when needed, for the many-many laughs, and for inspiring and motivating me to get finished. And the **Borrowash Buddies**; my support system for the past 25 years, for your friendship, for cheering me up during the difficult moments, and for being there to share the joyful ones. I am #blessed to have such wonderful women in my life, thank you all.

And to **Alec**, my heart, without your constant support and cooking capabilities, I wouldn't have made it. Thank you for putting up with me when I turned into a tired, hangry nightmare, and for being patient and kind always. Without you these past few years would have been boring and vegetable-less, and both me and my stomach are eternally grateful. I love you x.

Table of contents

List of tables

List of figures

1 INTRODUCTION

1.1 Thesis overview

Prognostic models provide personalised estimates of the risk of future events for patients in a given health state (Harrell, 2015, Steyerberg et al., 2013). Clinicians and patients may utilise prognostic models to predict the risks of future events, enhancing informed decision making and enabling tailored treatment strategies (Steyerberg et al., 2013). In order to enhance decision making, prognostic models should produce reliable (i.e. accurate) estimates of the absolute risk of the event. Ideally, prognostic models estimate the risk of a clinically relevant outcome, often through the development of a statistical equation, such as via multivariable regression techniques (Steyerberg et al., 2013). Time-to-event analysis (also known as survival analysis) methods are utilised when both the occurrence of an event of interest and the time until the occurrence are of interest (Harrell, 2015).

However, it is common in clinical practice to have multiple events that could occur, which preclude or alter the probability of the event of interest occurring (Koller et al., 2012). These so-called *competing* events, if not appropriately accounted for when developing prognostic models, can result in inflated absolute risk estimates (Andersen et al., 2012, Berry et al., 2010, Koller et al., 2012, Wolbers et al., 2009), referred to as "*competing risks bias*" (Schatzkin and Slud, 1989, Schumacher et al., 2016, Walraven and McAlister, 2016). For example, the risk of second hip fracture at 10 years was estimated to be 21% when ignoring the competing risk of death. This is 1.75 times greater than the true risk of 12%, estimated by appropriately accounting for deaths (Berry et al., 2010). Such biases in absolute risk predictions may result in inaccurate individual risk predictions which could impact the decisions and treatment strategies initiated by clinicians and their patients. This may be a particular issue when

developing prediction models in frail populations, such as the elderly, those with multimorbidities, or the critically ill, as in these settings individuals are more likely to experience a number of competing outcomes prior to the outcome of interest (Koller et al., 2012).

In order to avoid competing risks bias, the appropriate competing risks statistical methods, rather than standard time-to-event methods, should be applied. However, while the theory of competing risks has been around since the 1760s (Bernoulli, 1760), investigations have revealed that the methods are not being applied adequately. A systematic review of scientific articles published in high-impact clinical journals between 2007 and 2010 found competing risks bias was present in 67% (24/35) of eligible studies (Koller et al., 2012). Additionally, a recent systematic review of articles that reported Kaplan-Meier estimates, published in prominent medical journals in 2013, revealed almost half (46/100) of the studies were susceptible to competing risks bias (Walraven and McAlister, 2016). An update of this review, focusing on articles published in the New England Medical Journal in 2015, similarly found almost half (25/51) of studies were susceptible to this bias (Schumacher et al., 2016). Finally, a systematic review of reports of randomized controlled trials published in 2015 found 77.5% (31/40) of the studies were susceptible to competing risks bias (Austin and Fine, 2017). Thus, it is likely that competing risks bias is present in prediction model research, but this needs to be formally examined.

Though the potential impact of competing events on prognostic model research could be substantial, limited research has been conducted in this area. Therefore, the focus of this thesis is the investigation of the presence and impact of competing events on prognostic model research. This chapter lays the foundations on which the thesis is built. An overview of prognosis research is presented, followed by an introduction to statistical methods for standard time-to-event analysis and competing risks analysis.

Methods for developing and validating prognostic models are presented, and the aims of the thesis are outlined.

1.2 Prognosis research

Prognosis, a key concept in evidence based medicine, refers to the risks of future health outcomes in patients with a given disease or health condition (Hemingway et al., 2013). The focus of a number of clinical actions (including screening, diagnosis, and therapy) is to understand, predict, and improve individual prognoses (Steyerberg, 2008). Thus, prognosis research seeks to investigate the relations between future outcomes in patients with a given health state, in order to improve health (Hemingway et al., 2013). This contrasts with diagnostic research, which investigates the probability that a disease or condition is already present in a patient. Prognosis research aids understanding of the natural history of the health state, indicating the most likely course of progression or recovery, identifying factors associated with the risk[i] of future events, and identifying groups of individuals most at risk of these events (Steyerberg, 2008). The study of prognosis is crucial for the increased understanding and improvement of patient outcomes, and should thus be integral to clinical decision making and healthcare policy (Hemingway et al., 2013).

The PROGnosis RESearch Strategy (PROGRESS) Partnership[ii] published a series of articles focusing on each of four research themes, listed below, which together outline a framework for prognosis research (Hemingway et al., 2013):

[i] The term “risk” is used throughout the thesis to convey the probability of a future event. While the connotation is that events are negative, this is not always the case e.g. predicting disease recovery.
[ii]The PROGRESS Partnership is a UK Medical Research Council (MRC) funded, international, interdisciplinary collaboration developing understanding in research into quality of care outcomes, prognostic factors, risk prediction models, and predictors of differential treatment response.

1. **Overall prognosis research** aims to describe and explain the average prognosis of future outcomes in relation to current diagnostic and treatment practices (Hemingway et al., 2013).
2. **Prognostic factor research** aims to identify any measure (prognostic factor) that is associated with a subsequent clinical outcome among people with a particular disease or health condition (Riley et al., 2013).
3. **Prognostic model research** utilises combinations of multiple prognostic factors to predict the risk of future clinical outcomes in individual patients (Steyerberg et al., 2013).
4. **Stratified medicine research** involves tailoring therapeutic decisions to an individual or groups of individuals based on their predicted risk and/or predicted response to therapy (Hingorani et al., 2013).

This thesis focuses on PROGRESS theme three (prognostic model research) as methods for the development and validation of multivariable prognostic models for time-to-event outcomes are explored. Prognostic models may be referred to as clinical prediction models/tools/rules, prognostic indices/scores, prediction models, predictive scores, risk scores, scoring systems, risk stratification tools, amongst others. Ideally, prognostic models estimate absolute risk predictions of a clinically relevant outcome, derived through a formal combination of multiple prognostic factors, typically by applying multivariable regression techniques (Steyerberg et al., 2013). Prognostic model research can be applied to identify high risk groups for targeted screening or prevention interventions, to aid therapeutic decision making and follow-up strategies, and in further medical research[i] (Steyerberg, 2008). An example of a well-known prognostic model with a time-to-event outcome, used in daily clinical practice, is provided in Figure 1.1.

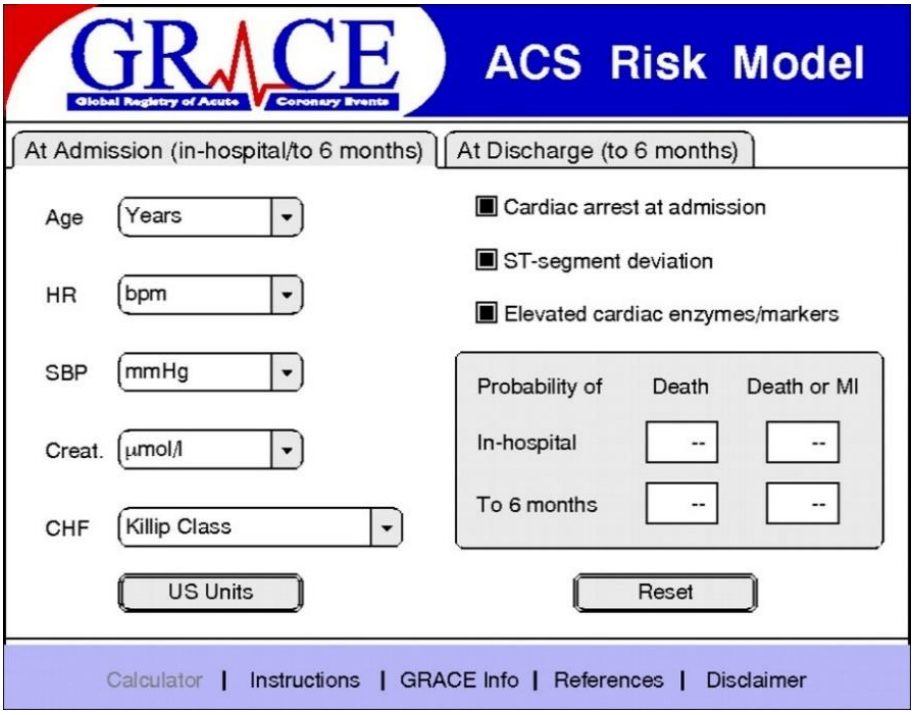
[i] Such as stratification in randomised controlled trials or adjusting baseline imbalance in observational studies.

Figure 1.1: Example of a prognostic model with time-to-event outcome - GRACE

The GRACE risk scores (Fox et al., 2006) are NICE[i] recommended (NICE CG94) prognostic models which estimate in-hospital and 6 month risks of death and myocardial infarction at admission and discharge for patients presenting with acute coronary syndrome. The models were developed using data from 21,688 patients included in a prospective multinational observational study.

Cox regression analysis was utilised to develop the models, which contain eight prognostic factors, including the patient's age, heart rate (HR), systolic blood pressure (SBP), and serum creatinine (Creat.) measures and whether the patient has congestive heart failure (CHF), cardiac arrest at admission, ST-segment deviation, or elevated cardiac enzymes/markers.

The GRACE risk scores were developed into an app, depicted below, and have been integrated into electronic medical record systems used in daily clinical management of acute coronary syndrome patients worldwide.



GRACE risk calculator (www.outcomes.org/grace) (Fox et al., 2006)[ii]

Numerous tutorials for prognostic model research have been published, including the BMJ[iii] series "Prognosis and prognostic research" (Moons et al., 2009b, Royston et al., 2009, Altman et al., 2009, Moons et al., 2009a), the "Risk prediction models" series (Moons et al., 2012b, Moons et al., 2012a), and the "Clinical prediction models" (Steyerberg, 2008) and "Prognosis Research in Healthcare: Concepts, Methods, and

[i]The National Institute for Health and Care Excellence (NICE) is a public body of the department of health in the UK, providing national guidance and advice to improve health and social care.
[ii] Reproduced from "Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE)" Fox et al., 333, 2006, with permission from BMJ Publishing Group Ltd.
[iii]The British Medical Journal (BMJ) is an international peer-reviewed medical journal.

Impact" (Riley et al., 2019a) books. Although the number of prognostic models in medical and health research have proliferated in recent years, current evidence indicates deficiencies in the statistical methods (Hemingway et al., 2009) and poor quality of reporting (Collins et al., 2015), prompting the development of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (Collins et al., 2015). The above texts discuss the principles and methods for prognostic model research, though there remains much debate regarding ideal practices. However, it is generally accepted that prognostic model research can be split into different stages; these can be categorised as model development (with or without internal validation); model validation (with or without updating); and subsequent evaluation of impact in clinical practice (Collins et al., 2015, Steyerberg et al., 2013).

It has been commented that prognostic models are more likely to be accepted and successfully implemented if shown to be clinically credible[i] (Wyatt and Altman, 1995), and show evidence of accuracy, generalisability, and effectiveness (Steyerberg et al., 2013). Further, a number of general principles for good practice are proposed for clinically useful prognostic models (Moons et al., 2012b, Steyerberg, 2008), outlined in Box 1.1. A recent review of prognostic models found models were more likely to be reliable when; developed using a large, high-quality dataset; based on a study protocol with a sound statistical analysis plan; validated in independent datasets obtained from different locations (Steyerberg et al. 2013).

[i] Clinically credible in this context refers to prognostic models which are supported by leading professionals and which contain all clinically important prognostic factors.

Box 1.1: Principles for good practice for clinically useful prognostic models

1. There should be a clear aim to predict a clinically meaningful outcome; which ought to be assessed objectively and without bias.
2. Participants should be representative and from a well-defined cohort.
3. A pre-determined selection of standardised, clinically relevant, and readily-available prognostic factors should be collected, at their intended moment of use (i.e. the time of prediction).
4. Follow-up time ought to be sufficient for the appropriate number of events to be observed, i.e. the sample size should be sufficient. In particular, it is often recommended there be at least ten events per candidate[i] prognostic factor (Peduzzi et al., 1995)
5. There should be an intention to apply the model in practice and/or research.

The remainder of this chapter introduces the statistical considerations for prognostic model research, specifically focusing on fundamental statistical concepts for time-to-event outcomes, including the presence of competing events, and considerations for the development and validation of prognostic models with time-to-event outcomes. Then from Section 1.7, the aims and objectives of the thesis are outlined.

1.3 Statistical analysis of time-to-event outcomes

Prognostic models predict the risks of an outcome within a specific time period (Steyerberg et al., 2013). When both the occurrence and the time until the occurrence of an event are of interest (such as time to next epileptic episode), or when long-term outcomes are of interest (such as 10 year risk of cardiovascular disease), the optimal approach is to apply time-to-event (also known as survival) analysis techniques (Harrell, 2015). Time-to-event analysis examines the time between a suitable starting point and the occurrence of an event. In prognosis research the starting point is usually patient diagnosis, first presentation of the health condition in clinical practice, or entry into a given health state, and the event is clinically relevant and binary (Collins et al.,

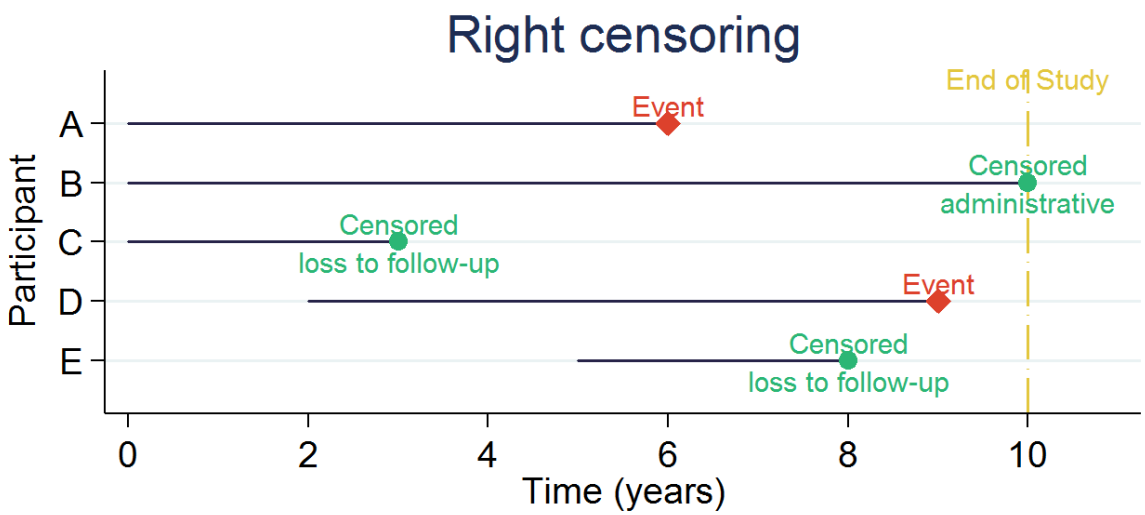
[i] Candidate prognostic factors refers to the group of prognostic factors considered for use in the prognostic model Thus may include prognostic factors which are not included in the final prognostic model due to the application of a selection process (see Section 1.5.2).

2015). The response variable, the time to the event, may have an unusual distribution; it is usually continuous but is restricted to be positive, thus often has a skewed (not normal) distribution. Additionally, it is unlikely that the event of interest will be observed for all participants; some participants may not experience the event during the study and some may be lost to follow-up. Time-to-event analysis can incorporate incompletely determined outcome information (censoring), which is likely to occur with long-term outcomes, thus it uses data more efficiently than other regression methods (such as logistic regression) (Harrell, 2015).

1.3.1 Censoring

Censoring occurs when a participant is not observed to experience the event of interest during the study follow-up period. There are a number of censoring mechanisms which may result in the reporting of censored observations, including left, right, and interval censoring (Harrell, 2015). Left censoring occurs when a participant experiences the event prior to the study start time. Interval censoring occurs when the event is known to have occurred during a time interval though the exact time is unknown, such as when an outcome is assessed at periodic clinical examination. Both left and interval censoring are not often present in prognostic model research and are thus not considered further in this thesis. Right censoring may occur if participants have not experienced the event by the end of follow-up (administrative censoring), or if participants stop being observed before the end of the study (loss to follow-up). A graphical representation of time-to-event data containing right-censoring is provided in Figure 1.2.

Figure 1.2: Example of time-to-event data with right censoring.



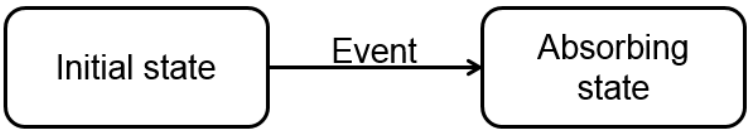
Participant follow-up time is shown by the solid blue lines.
Participant A experiences the event of interest after six years.
Participant B does not experience the event of interest during the study period, so is right-censored and contributes ten years of event-free time to the analysis.
Participant C is lost to follow-up, so is right-censored; and contributes three years of event-free time to the analysis.
Participant D joins the study at two years, and experiences the event at nine years; their time-to-event is 9-2 = 7 years.
Participant E joins the study at five years, and withdraws from the study at eight years, so is right-censored and contributes 8-5 = 3 years of event-free time to the analysis.

A key assumption of standard time-to-event analysis is non-informative (or independent) censoring; that those participants who are censored have the same probability of experiencing the event as others still at risk at the time the censoring occurred (Geskus, 2015). Under this assumption, removing those who are censored from the analysis at the time of censoring does not affect the overall risk of an event going forward. If informative censoring exists, then the time-to-event estimates may be biased. For example, if participants who are lost to follow-up withdrew from the study because they felt particularly unwell, their risk of experiencing an adverse event might be comparably worse than those who remain in the study. Whether the assumption of non-informative censoring is reasonable in the presence of other events, particularly ones which can alter the risk of the event of interest (competing events), will be discussed in detail later in this chapter.

1.3.2 Introduction of the multi-state structure

Time-to-event analysis methods are now described in the context of a multi-state modelling structure. Under this structure, participants transition between states following the occurrence of an event. The simplest multi-state structure is depicted in Figure 1.3, and begins with all participants in a single *initial* state. Participants remain in this state until the occurrence of an event of interest, at which time they transition into a second state. Participants then remain in the second state until the end of the study, as there is no chance of exiting the second state it is referred to as an *absorbing* state. This structure can be used to describe standard time-to-event analysis techniques; for example, when analysing survival, all participants are alive (the initial state) until death (the event) after which they are dead (the absorbing state).

Figure 1.3: Simple multi-state structure.



Boxes represent states and arrows represent transitions.
All participants begin in the initial state. Following an event, participants transition into the absorbing state.

1.3.3 Key functions in time-to-event analysis

Time-to-event analysis comprises a number of functions which are of interest in prognostic model research. These measures are detailed below for a continuous non-negative random variable T, which represents the time between a defined starting point $t = 0$ and the occurrence of an event of interest. Hence, T has a probability distribution with an underlying probability density function $f(t)$.

1.3.3.1 The survival and cumulative incidence functions

In prognostic model research, the key outcome of interest is the risk of either experiencing or surviving the event of interest before a given time t, for example the risk of relapse in the five years following breast cancer surgery. It is common in time-to-event analysis to define a statistical model using the survival function $S(t)$, a time

dependent function which estimates the probability of an individual not experiencing (i.e. surviving) the event of interest prior to a given time t, written:

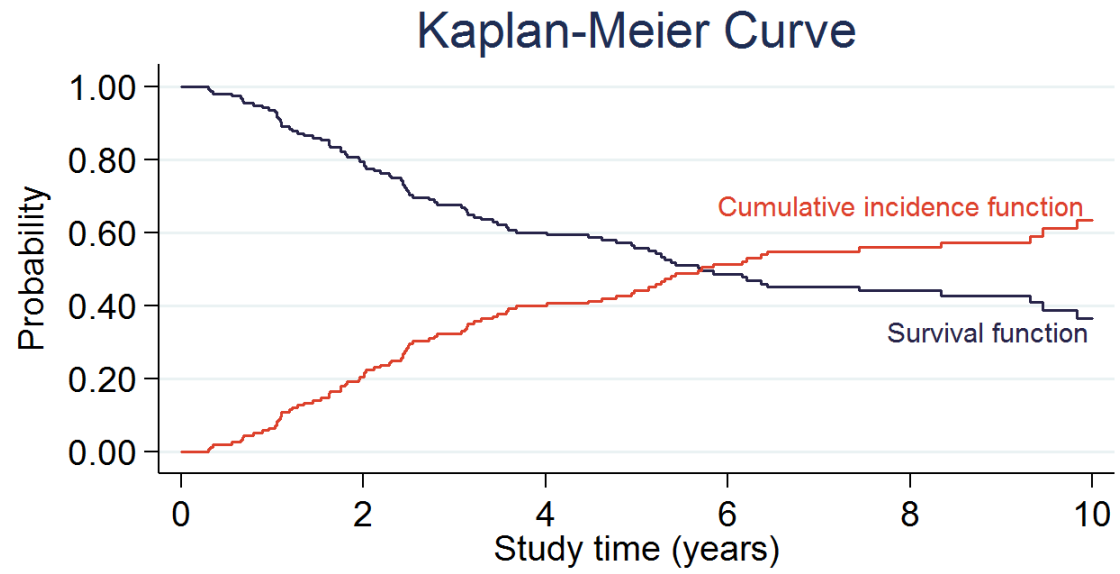
$$S(t) = P(T \geq t), \quad 0 < t < \infty \qquad \textit{Equation 1.1}$$

However, in prognostic model research, it is often more intuitive to communicate risk as the probability of an individual experiencing the event prior to time t (such as five-year mortality risk). This is estimated using the cumulative incidence function $F(t)$, written:

$$F(t) = P(T < t) = 1 - S(t), \quad 0 < t < \infty \qquad \textit{Equation 1.2}$$

The survival function, and thus the cumulative incidence function, can be estimated non-parametrically using the Kaplan-Meier method (Kaplan and Meier, 1958). This method estimates the probability of survival at each unique event time, producing a step-function which is often shown graphically as a Kaplan-Meier curve, an example of which is provided in Figure 1.4.

Figure 1.4: Example of a Kaplan-Meier curve.



A Kaplan-Meier curve depicting survival and cumulative incidence functions for disease recurrence and death using a selection of patients from the Rotterdam breast cancer study data (Foekens et al., 2000).
The Kaplan-Meier curve commonly depicts the probability of an event not occuring over time, i.e. the survival function (blue line). A curve depicting the risk of an event occuring over time, i.e. the cumulative incidence function (red line), is often referred to as the 1 – Kaplan-Meier curve.

At the study start time, i.e. when $t = 0$, all of the participants are event free, thus $S(t = 0) = 1$ and $F(t = 0) = 0$. The probability of experiencing the event (cumulative incidence) can only increase over time as more events occur, thus the probability of not experiencing the event (survival) can only decrease over time.

1.3.3.2 The hazard and cumulative hazard functions

The rate at which these survival and cumulative incidence functions change over time is also of interest in prognostic model research. This is measured using the hazard function $h(t)$, which estimates the instantaneous rate of experiencing the event of interest at time t, conditional on the individual being at risk at that time (i.e. having not experienced the event prior to time t). The hazard function over time, for a small time interval δ, is written:

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta}, \ \delta > 0 \qquad \textbf{\textit{Equation 1.3}}$$

Another related quantity is the cumulative hazard function $H(t)$, which estimates the total amount of hazard experienced up until time t, written:

$$H(t) = \int_0^t h(u)\, du \qquad \textbf{\textit{Equation 1.4}}$$

1.3.3.3 Theoretical relationships between key functions

In time-to-event analysis, there exists a direct relationship between the risk (survival and cumulative incidence functions) and rate (hazard and cumulative hazard functions) of an event. For example, the survival function can be written as a transformation of the hazard function, and vice versa:

$$S(t) = \exp\left(-\int_0^t h(u)\, du\right), \quad h(t) = -\frac{d}{dt} \ln[S(t)] \qquad \textbf{\textit{Equation 1.5}}$$

Due to the relationship between the survival and cumulative incidence functions (Equation 1.2), the above equations can also be written:

$$F(t) = 1 - \exp\left(-\int_0^t h(u)\,du\right), \quad h(t) = -\frac{d}{dt}\ln[1 - F(t)]$$ ***Equation 1.6***

Similarly, due to the relationships between the hazard and the cumulative hazard function (Equation 1.4), these equations can also be written:

$$S(t) = \exp\big(-H(t)\big), \quad H(t) = -\ln[S(t)]$$ ***Equation 1.7***

These relationships result in what is often referred to as a *one-to-one relationship* between the risk and rate of the event of interest.

1.3.4 Time-to-event regression

Prognostic models usually need to incorporate combinations of multiple prognostic factors into the statistical model to accurately predict the risk of future outcomes (Steyerberg et al., 2013). Regression methods allow for multiple independent prognostic factors (variables) to be analysed simultaneously within a single *multivariable* model (Harrell, 2015). The Cox proportional hazard model (Cox, 1972a) is considered to be the most commonly used model for the analysis of time-to-event data (Royston and Lambert, 2011). Indeed, it has been shown to be one of the most frequently used regression models in prognostic model research, alongside logistic regression (Collins et al., 2015, Royston et al., 2009, Snell, 2015). However, recent research has highlighted many statistical advantages of using flexible parametric models, specifically the Royston-Parmar model (Royston and Parmar, 2002), rather than Cox models in prognostic model research (Snell, 2015). These two models are discussed in greater detail below for the incorporation of $\mathbf{X_i} = (x_1, x_2, \ldots)^T$, a vector of prognostic factors for participant i.

1.3.4.1 Cox proportional hazards models

The Cox proportional hazards model estimates regression coefficients on the hazards scale:

$$h_i(t) = h_0(t)\exp\big(\boldsymbol{\beta}^T\mathbf{X_i}\big)$$ ***Equation 1.8***

In which $h_0(t)$ represents the baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ represents a vector of regression coefficient estimates. The Cox model estimates the regression coefficients $\boldsymbol{\beta}$ (also referred to as log hazard ratios) by maximising the partial likelihood on the hazards scale (Cox, 1972a). Following an exponential transformation, these estimates represent hazard ratios, which reflect the relative change in hazard rate for a unit increase in the associated prognostic factor. The baseline hazard function corresponds to the hazard rate over time when all prognostic factors are equal to 0. To make the interpretation of this function more meaningful, prognostic factors are often centred about their mean value prior to estimation, or similarly the linear predictor $\boldsymbol{\beta}^T\mathbf{X_i}$ (also known as prognostic index) is centred (Royston and Altman, 2013). The hazard ratio estimates obtained from the Cox model represent a change in hazard rate from a hypothetical "baseline participant". If prognostic factors are centred, this "baseline participant" will be more representative of the participants, rather than, for example, a participant aged 0. The relative estimates produced by the model are useful for identifying differences in risks between participants with different prognostic characteristics, thus are essential for prognostic factor research. An example of a well-known prognostic model developed using Cox regression is provided in Box 1.4.

The Cox model does not make any distributional assumptions with regard to the shape of the baseline hazard function $h_0(t)$, thus is described as a *semi-parametric* model. As a consequence, the model only estimates hazard ratios and does not directly estimate absolute risks (Royston and Altman, 2013), the more intuitive and interpretable output for prognostic model research.

Box 1.2: Example of a prognostic model developed using Cox regression

The Nottingham Prognostic Index (Haybittle et al., 1982) is a well-known prognostic model which estimates the five year risk of death in patients that had a primary operable breast cancer. The model was developed using retrospective data from 387 patients from the Nottingham Breast Cancer Study.

Cox regression analysis was utilised to develop the model, which contains three prognostic factors; tumour **size** (cms), lymph-node **stage** (coded A=1, B=2, C=3), and tumour **grade** (coded I=1, II=2, III=3). The model is described using the following linear predictor:

$$\mathbf{I} = (\mathbf{0.17} \times \mathbf{size}) + (\mathbf{0.76} \times \mathbf{stage}) + (\mathbf{0.82} \times \mathbf{grade})$$

The individual regression coefficients reflect the relationship between the prognostic factor and the five year risk of death. A 1cm increase in tumour **size** results in a hazard ratio equal to:

$$\mathbf{HR} = \mathbf{exp}(\mathbf{0.17}) = \mathbf{1.19}$$

The hazard of death increases by a factor of 1.19 for each 1cm increase in tumour size.

The index may be used to estimate the risk of death for a participant with a **grade II** tumour (coded II=2) of **size 1.5**cms and lymph-node **stage A** (coded A=1) as follows:

$$\mathbf{I} = (\mathbf{0.17} \times \mathbf{1.5}) + (\mathbf{0.76} \times \mathbf{1}) + (\mathbf{0.82} \times \mathbf{2}) = \mathbf{2.66}$$

This participants hazard of five year death is obtained following an exponential transformation of the linear predictor:

$$\mathbf{HR} = \mathbf{exp}(\mathbf{2.66}) = \mathbf{14.30}$$

Hence this participant has a 14.3 times increased hazard in comparison to the "baseline participant".

The baseline participant has a tumour of size 0cms with implausible grade and lymph-node stage classifications (as there exists no categories for grade=0 and stage=0)[i].

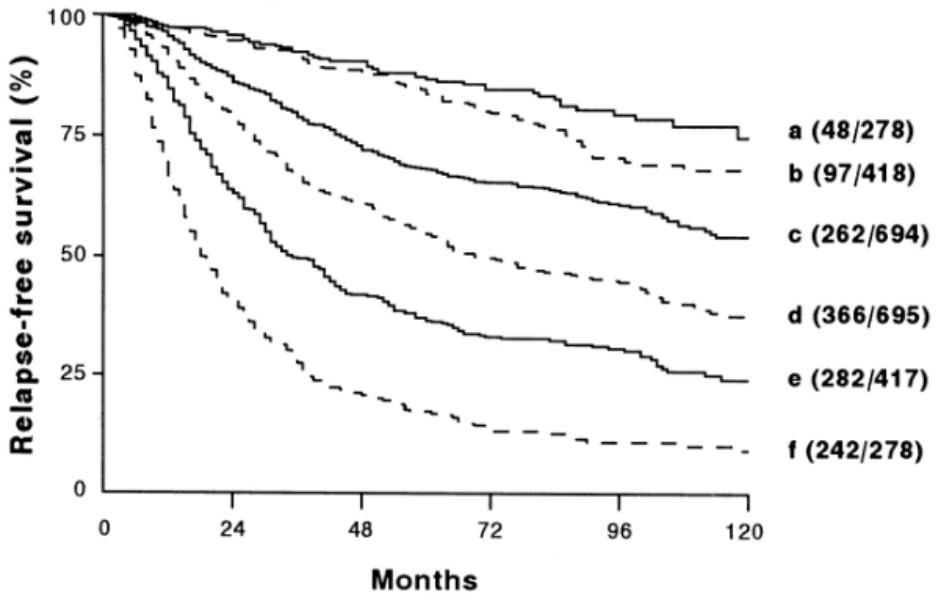
A key assumption made by the Cox (and other time-to-event) models is of proportional hazards (Harrell, 2015), that is the hazard ratios remain constant over time. This assumption may be examined using log-log plots (Royston and Lambert, 2011) or applying a Grambsch-Theneau test (Grambsch and Therneau, 1994) of scaled Schoenfeld residuals (Schoenfeld, 1982). The proportional hazards assumption may be relaxed through the incorporation of interaction terms between prognostic factors and time (Cox, 1972b).

In prognostic model research, estimates of the absolute risk of an event occurring over time are more intuitive and easier to communicate to clinicians and patients. A

[i]It is not plausible for a *real* patient to have a tumour of size 0cms, and not providing baseline categories for the other two categories causes further problems. The interpretation of the estimated relative risks become complicated, thus highlighting the importance of centring prognostic factor values prior to model development.

commonly used approach to obtain absolute risk estimates from a prognostic model developed using Cox regression is through the formation of risk groups (Royston and Altman, 2013). The linear predictor $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{\mathbf{i}}$ derived from the Cox model is split to create the risk groups. Kaplan-Meier curves for each risk group can then be utilised to obtain absolute risk predictions for each group at given time points, as depicted in Figure 1.5.

Figure 1.5: Example of Kaplan-Meier curves for risk groups.



Kaplan-Meier curves showing two year relapse-free survival for 2,780 patients with primary, operable, invasive breast cancer (Foekens et al., 2000)[i].
Risk groups were created by splitting the Cox model linear predictor at the 10, 25, 50, 75, and 90 perecentiles, producing groups a-f. The numbers in parentheses refer to the number of events/ patients per group.

The disadvantage of grouping participants into risk groups is that information from the prognostic model is lost (Harrell, 2015) and individual predictions are less accurate (Steyerberg et al., 2013). An alternative and preferable approach is to utilise the relationship between the hazard, cumulative hazard, survival, and cumulative incidence functions discussed previously to determine an equation for the absolute risk of an event occurring over time, i.e. the cumulative incidence function:

$$\mathrm{F_i(t)} = 1 - \mathrm{S_0(t)}^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{\mathbf{i}})} \qquad \textit{Equation 1.9}$$

[i] Adapted from Cancer Research, 2000, 60/3, 636-643, Foekens et al. "The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients", with permission from AACR.

Where $S_0(t) = \exp(-H_0(t))$ represents the baseline survival function and $H_0(t) = \int_0^t h_0(u)\,du$ represents the baseline cumulative hazard function.

Though not directly estimated in Cox regression, it is possible to derive a non-parametric estimate for $H_0(t)$, known as Breslow's estimate (Breslow, 1972). Let R_j denote the set of study participants at event time t_j and l represent those at risk (i.e. the risk set). Assuming no tied event times, and utilising the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ obtained from the Cox regression; the baseline cumulative hazard function can be approximated using the following formula:

$$\widehat{H_0}(t) = \sum_{j:t_j \le t} \frac{1}{\sum_{l \in R_j} \exp(\boldsymbol{\beta}^T \mathbf{X_l})} \qquad \textit{Equation 1.10}$$

The above formulae estimates $\widehat{H_0}(t)$ at each unique event time, producing a step-function. The estimates can be incorporated into Equation 1.9 and transformed to provide individual absolute risk estimates for the event of interest. Though the above methods are available to develop prognostic models which give absolute risk predictions, models built using Cox regression are often reported without estimation of the baseline hazard function (Royston and Altman, 2013). This limits the application of the prognostic models, as only estimates of relative risks (and not absolute risks) can be calculated.

1.3.4.2 Royston-Parmar flexible parametric models

Parametric models are fully specified, i.e. they make distributional assumptions about the baseline hazard functions. As such the models can directly estimate the absolute risks of events occurring over time, making them ideal for prognostic model research. Common examples of parametric models include the exponential model (Equation 1.11), which assumes a constant (λ) hazard over time, and the Weibull model (Equation 1.12), which assumes a monotone increasing or decreasing hazard over time:

$$h_i(t) = \lambda \exp\left(\boldsymbol{\beta}^T \mathbf{X_i}\right) \qquad \textit{Equation 1.11}$$

$$h_i(t) = \lambda \gamma t^{\gamma-1} \exp\left(\boldsymbol{\beta}^T \mathbf{X_i}\right) \qquad \textit{Equation 1.12}$$

Though explicitly defined, the shape of the baseline hazard functions in these models are limited and often not considered to be clinically plausible. Flexible parametric models, however, utilise spline functions to obtain a smooth estimate of the observed baseline hazard function (Royston and Lambert, 2011). These models can capture more flexible and clinically plausible baseline hazard functions than standard parametric models. Royston-Parmar models (Royston and Parmar, 2002) explicitly model the baseline cumulative hazard function with a large degree of flexibility, by utilising restricted cubic splines.

Restricted cubic splines are piecewise cubic functions which join at predefined values of t; these points are called knots. Continuity constraints, which force the functions to have continuous first and second derivatives, ensure the splines connect smoothly at these knots. The functions before the first and after the final knot are restricted to be linear (Harrell, 2015). When modelling the time to an event, the first knot η_{min} is defined at the time to the first observed (uncensored) event, and the final knot η_{max} is defined at the time of the last observed event, these are the boundary knots (Royston and Lambert, 2011). A restricted cubic spline function for $\ln(t)$, with M interior knots, a vector of knot locations $\mathbf{N} = (\eta_{min}, \eta_1, \dots, \eta_M, \eta_{max})$, and parameters $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_{M+1})$, is written:

$$\text{spline}\{\ln[t] | \boldsymbol{\gamma}, \mathbf{N}\} = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{M+1} z_{M+1} \qquad \textit{Equation 1.13}$$

In which the parameters γ_j are estimated and the variables z_j are derived as follows:

$$z_1 = \ln[t]$$

$$z_j = \left(\ln[t] - \eta_j\right)_+^3 - \phi_j (\ln[t] - \eta_{min})_+^3 - \left(1 - \phi_j\right)(\ln[t] - \eta_{max})_+^3 \qquad \textit{Equation 1.14}$$

where $\phi_j = {}^{\eta_{max} - \eta_j}/_{\eta_{max} - \eta_{min}}$,

for $j = 2, \ldots, M + 1$ and $u_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{if } u \leq 0 \end{cases}$ (Royston and Lambert, 2011).

A restricted cubic spline function with M internal knots is estimated with $M + 1$ degrees of freedom. The number of knots required to capture the shape of the hazard function can be determined through visual inspection of graphs depicting the different spline functions or through the comparison of Akaike and Bayesian information criterion estimates (AIC and BIC) (Royston and Lambert, 2011). However, functions with more than five internal knots are seldom required (Harrell, 2015), and functions containing between one and four internal knots have been advised specifically for modelling the baseline hazard function (Royston and Lambert, 2011). The position of the knots is usually determined using centiles of uncensored event times (Royston and Lambert, 2011). Numerous sensitivity analyses in various applications have shown that, once a sensible number of knots have been determined, any results are fairly robust to changes in knot locations (Hinchliffe, 2013).

The Royston-Parmar flexible parametric model estimates regression coefficients through maximum likelihood on the more stable log cumulative hazard scale:

$$\ln[H_i(t)] = \ln[H_0(t)] + \boldsymbol{\beta}^T \mathbf{X_i} \qquad \textbf{\textit{Equation 1.15}}$$

The log baseline cumulative hazard function $\ln[H_0(t)]$ is modelled as a smooth non-linear function of log time using restricted cubic splines:

$$\ln[H_i(t)] = \text{Spline}\{\ln[t]\} + \boldsymbol{\beta}^T \mathbf{X_i} \qquad \textbf{\textit{Equation 1.16}}$$

Again in prognostic model research, absolute risks, estimated using the survival and cumulative incidence functions, are of greater interest. Utilising the relationship between the survival and cumulative hazard functions (Equation 1.7) it is possible to determine the survival function for the Royston-Parmar model:

$$\ln[S_i(t)] = -\exp\left(\text{spline}\{\ln[t]\} + \boldsymbol{\beta}^T \mathbf{X_i}\right) \qquad \textbf{\textit{Equation 1.17}}$$

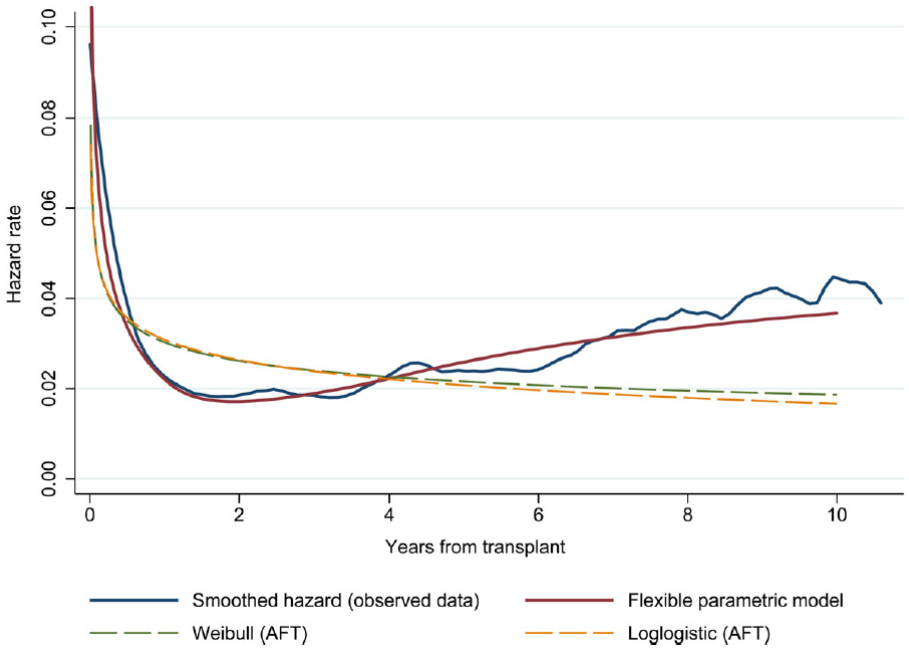
An example of a prognostic model developed using Royston-Parmar flexible parametric regression methods is provided in Box 1.3.

Box 1.3: Example of a Royston-Parmar flexible parametric prognostic model

A prognostic model to predict post-transplant survival for patients following kidney transplantation was developed using data from 12,000 recipients of kidney transplants from the UK Transplant Registry (Li et al., 2016).

A flexible parametric proportional hazards model was utilised to develop the model, which contains six prognostic factors; recipient age, recipient gender, pre-emptive transplant, primary renal diagnosis, donor hypertension, and donor age.

The number of knots required for the restricted cubic spline characterising the baseline hazard function was determined through comparisons of AIC. The model with optimal fit contained two internal knots (3 d.f.), depicted using the red line in the graph below, knot locations were not discussed:



Comparison of smooth hazard functions to observed data (Li et al., 2016)[i]

The regression coefficients for the spline function are provided below:

Baseline hazard (log hazard scale)	Coefficient	p-value	95% CI
Restricted cubic spline 1	1.03	<0.001	0.97 – 1.09
Restricted cubic spline 2	-0.08	0.001	-0.12 - -0.03
Restricted cubic spline 3	-0.14	<0.001	-0.16 - -0.12
Constant	-3.97	<0.001	-4.31 - -3.63

Coefficients for baseline hazard function spline terms, adapted from (Li et al., 2016)[i]

If the baseline cumulative hazard function is modelled with appropriate degrees of freedom, the regression coefficients estimated with a Royston-Parmar model are very

[i] "Predicting patient survival after deceased donor kidney transplantation using flexible parametric modelling", Li et al., BMC Nephrology, 2016 17:51, CC BY 4.0.

similar to those estimated with a Cox model (Snell, 2015). The Royston-Parmar model described above assumes proportional hazards. The proportional hazards assumption can be relaxed through the incorporation of interaction terms between prognostic factors and the spline functions for $\ln(t)$. Alternative Royston-Parmar models include proportional odds (a generalisation of log-logistic models) and probit (a generalisation of log-normal models) models (Royston and Parmar, 2002), however these are not covered within this thesis.

There exist many statistical advantages for using Royston-Parmar models, rather than Cox models (Snell, 2015), these include:

1. The direct estimation of the baseline hazard function;
2. Practically identical regression coefficient estimates to Cox models;
3. The ability to graphically represent population-averaged survival curves;
4. The estimation of differences in absolute survival over time;
5. Deriving absolute risk predictions for individuals;
6. Identifying clinically important hazard ratios in relation to the baseline risk.

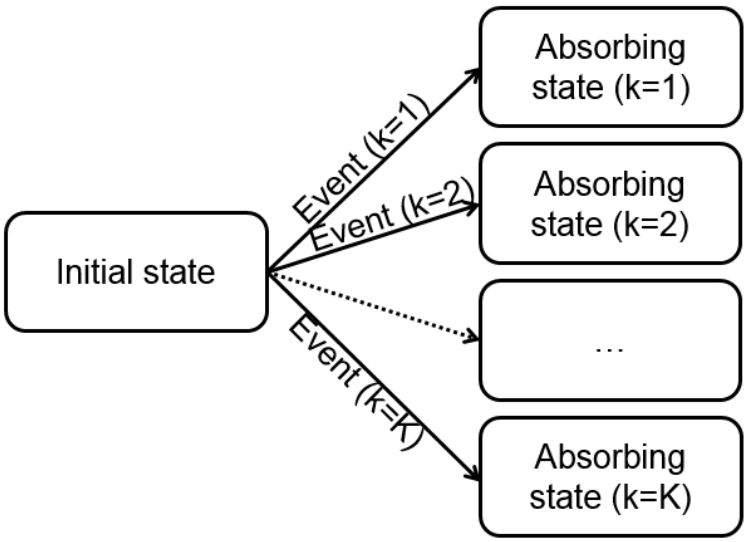
1.4 Statistical analysis of competing events

Competing risks are present when participants may experience an event other than the event-of-interest, and the occurrence of this *competing event* alters the probability of the event-of-interest occurring (Koller et al., 2012). The methods to appropriately account for competing events have been traced to a study investigating the consequences of mandatory vaccination against smallpox and the effects on population survival (Bernoulli, 1760). The standard time-to-event analysis methods described previously have been adapted to incorporate competing events.

1.4.1 The multi-state structure with competing events

The simple multi-state structure depicted in Figure 1.3 can be extended to incorporate multiple absorbing states, allowing participants to experience one of multiple (K) mutually exclusive *competing* events Figure 1.6.

Figure 1.6: Competing risks multi-state structure



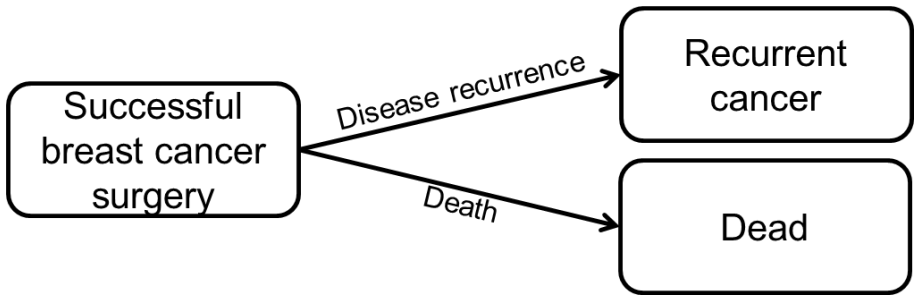
Boxes represent states and arrows represent transitions.
All participants begin in the initial state. The occurrence of an event of interest causes participants to transition into the associated absorbing state. Participants who experience competing events transition into a different absorbing state.

This structure can be used to describe competing risks analysis techniques; for example, when analysing a specific cause of death (such as cardiovascular deaths) in which other causes of death are competing events. All participants are alive (the initial state) until death from one cause or another (the events) after which they enter a cause-specific state (the absorbing states). When analysing a specific cause of death (such as cardiovascular death), other causes of death may be examined separately (i.e. have separate absorbing states), or can be combined into one competing state, namely other causes of death, such that $\mathrm{K} = 2$.

All-cause mortality (death) is generally considered a competing event for all non-fatal events, particularly when studying elderly or frail populations as death prevents any non-fatal event from occurring. The analysis of non-fatal events can be represented using a simple competing risks multistate structure with $\mathrm{K} = 2$ absorbing

states; for example, when investigating disease recurrence following breast cancer surgery (Figure 1.7).

Figure 1.7: Competing risks multi-state structure with K=2



Boxes represent states and arrows represent transitions.
All participants begin in the initial state (e.g. successful breast cancer surgery).
Following an event (e.g. disease recurrence, k=1), participants transition into the associated state (e.g recurrent cancer).
Participants who experience competing events (e.g. death) transition into a different absorbing state (e.g. dead).

All participants surviving surgery remain in the initial state until transition to another state occurs. For example, they may experience a disease recurrence (the event of interest, $\mathrm{k} = 1$) which instigates a transition into the recurrent cancer state[i]. Or they may die before disease recurrence (the competing event), and so transition into the other absorbing state (dead, $\mathrm{k} = 2$) and are no-longer at risk of experiencing disease recurrence (the event of interest). This example (breast cancer recurrence and the competing risk of death) will be used to introduce the functions of interest in prognostic model research, as the research often focuses on analysing a non-fatal event of interest with death as a competing event.

1.4.2 Key functions in time-to-event analysis with competing events

Similar to standard time-to-event analysis, competing risks analysis comprises a number of key functions which are of interest for prognostic model research. These key functions are similar to those obtained from a standard time-to-event analysis, but have been extended into the competing risks setting. These measures are detailed

[i] In this instance the "recurrent cancer" state is not absorbing; a patient may die following cancer recurrence, which would trigger a transition from the "recurrent cancer" state to the "dead" state. However, events which occur after the event of interest (cancer recurrence, $k = 1$) are often not of interest, in which case the state is considered an absorbing state.

below for a continuous non-negative random variable T, which represents the time between a defined starting point $t = 0$ and the occurrence of $\mathbf{k} = (1,2)$ mutually exclusive events. Thus, T has a probability distribution with an underlying probability density function $f(t)$.

1.4.2.1 All-cause functions

In prognosis research scenarios where multiple types of events can occur over time, the specific type of event is not always of importance or of interest. Rather, the research may focus on a combined (composite) outcome of all events (or causes). For example, when investigating recurrence-free survival following breast cancer surgery, it is not necessary to consider the events (disease recurrence and death) separately; instead the more relevant approach would be to investigate the occurrence of *either* disease recurrence *or* death. The rate of either event occurring is investigated using all-cause functions, which are equivalent to functions in standard time-to-event analysis (detailed in Section 1.3.3), with all mutually exclusive events combined into one composite event (i.e. all events transition into a single state) (Geskus, 2015).

1.4.2.2 Cause-specific functions

The cause-specific functions are relevant when the focus of the research is a specific event (referred to as the event of interest, $k = 1$), in the presence of competing events. In these instances, it is necessary to distinguish between the different types of events, and so allow the risk and rate of each event of interest to be investigated independently. For example, consider predicting the risk of disease recurrence (the event of interest) following breast cancer surgery; here it is important to account for the competing event of death, but predicting the actual risk of death in this instance is not of interest. The cause-specific functions provide relevant information about a specific event of interest ($k = 1$) from the K mutually exclusive events. Though the cause-specific functions are similar to those used in standard time-to-event analysis, a

number of subtle differences distinguish between the interpretation of the standard time-to-event and competing risks approaches.

The cause-specific cumulative incidence function

In the presence of competing events, the key measure of interest for prognostic model research is the risk of a participant experiencing the event of interest $(k = 1)$ prior to time t, allowing for the presence of competing events. For example, the five-year risk of disease recurrence following breast cancer surgery, acknowledging that some participants may die prior to relapse. This is the estimated the cause-specific cumulative incidence function $F_k(t)$, which is written:

$$F_k(t) = P(T \leq t | event = k) \qquad \textbf{\textit{Equation 1.18}}$$

The sum of all K cause-specific cumulative incidence functions is equal to the all-cause cumulative incidence function:

$$F(t) = P(T \leq t) = \sum_{k=1}^{K} P(T \leq t | event = k) = \sum_{k=1}^{K} F_k(t) \qquad \textbf{\textit{Equation 1.19}}$$

The cause-specific cumulative incidence function is not a proper probability distribution function, as it does not increase to one over time. This is because the competing events prevent the event of interest from occurring (Geskus, 2015). For example, it is possible for some participants to die prior to disease recurrence; these participants will no longer be at risk of recurrence, and thus the probability of recurrence will never reach 100%.

The cause-specific hazard function

Another measure of interest in prognostic research is the cause-specific hazard function, determined as the rate at which the cause-specific cumulative incidence function for the event of interest $(k = 1)$ changes over time. This measure is conditional on participants not experiencing any of the K mutually exclusive events prior to that time. The cause-specific hazard function $h_k(t)$ is written:

$$h_k(t) = \lim_{\delta \to 0} \left\{ \frac{P(t \leq T \leq t + \delta, \text{event} = k | T \geq t)}{\delta} \right\}$$ ***Equation 1.20***

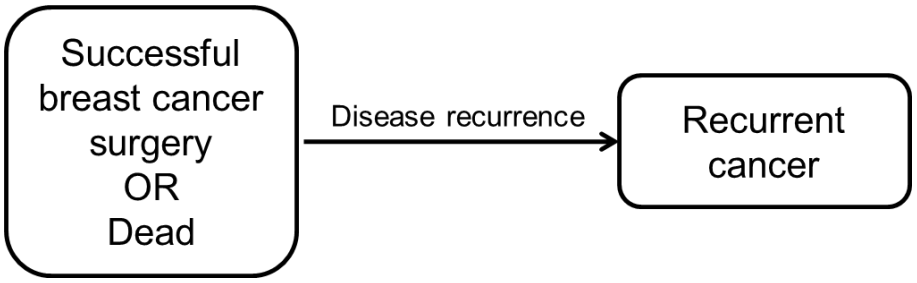
The participants that have experienced a competing event prior to time t transition out of the initial state at the time the competing event occurs, and are thus censored for the event of interest $(k = 1)$. The cause-specific hazard function is equal to the hazard function calculated using standard time-to-event analysis methods; competing events are not anticipated in a standard time-to-event analysis, thus are typically censored when they occur.

The cause-specific hazard ratio reveals the effect of a prognostic factor directly on the event of interest. It ignores (censors) other types of events that can occur (competing events), and hence describes the effect of a prognostic factor assuming the competing events cannot occur (Koller et al., 2012). Thus cause-specific hazard ratios resemble those derived using standard time-to-event analysis methods. These direct effects are important for answering etiological questions, where the focus is more on causality of a factor, and are typically more of interest in prognostic factor research.

1.4.2.3 Subdistribution hazard function

Another measure of interest in prediction model research is the subdistribution hazard, determined as the instantaneous rate at which the cause-specific cumulative incidence function for the event of interest $(k = 1)$ changes over time, conditional on participants not experiencing the event of interest prior to that time. To obtain these estimates the multi-state structure is modified (Figure 1.8).

Figure 1.8: Competing risks multi-state structure for subdistribution hazards



Boxes represent states and arrows represent transitions.
All participants begin in the initial state (e.g. successful breast cancer surgery).
Following the event of interest (e.g. disease recurrence, k=1), participants transition into the associated absorbing state (e.g recurrent cancer).
Participants who experience competing events (e.g. death) remain in the initial state. As these participants are not at risk of experiencing the event of interest, they remain in this initial state until the end of the study.

All participants surviving surgery remain in the initial state until disease recurrence (the event of interest, $\mathrm{k} = 1$), which instigates a transition into the recurrent cancer state. Those participants who die before disease recurrence (the competing event) remain in the initial state, but are no-longer at risk of experiencing disease recurrence, thus remain there until their known but unobserved censoring time (i.e. the end of the study) (Fine and Gray, 1999). The subdistribution hazard function $\lambda_{\mathrm{k}}(\mathrm{t})$, is written:

$$\lambda_{\mathrm{k}}(\mathrm{t}) = \lim_{\delta \to 0} \left\{ \frac{\mathrm{P}\left(\mathrm{t} \leq \mathrm{T} \leq \mathrm{t} + \delta, \mathrm{event} = \mathrm{k} \middle| \mathrm{T} \geq \mathrm{t} \cup (\mathrm{T} \leq \mathrm{t} \cap \mathrm{event} \neq \mathrm{k})\right)}{\delta} \right\} \qquad \textit{Equation 1.21}$$

The subdistribution hazard is useful for considering prognostic model research questions, where estimation of the absolute risk is of primary importance (Koller et al., 2012). The subdistribution hazard estimates the “real world” change in the risk of the event of interest, as it incorporates the indirect effects of a prognostic factor on the competing event. These indirect effects are important in prognostic model research as they explain the rate at which the risk of an event changes over time.

1.4.2.4 Summary of the functions of interest in competing risks analyses

The functions of interest are summarised below in Table 1.1.

Table 1.1: Descriptions of competing risks functions of interest

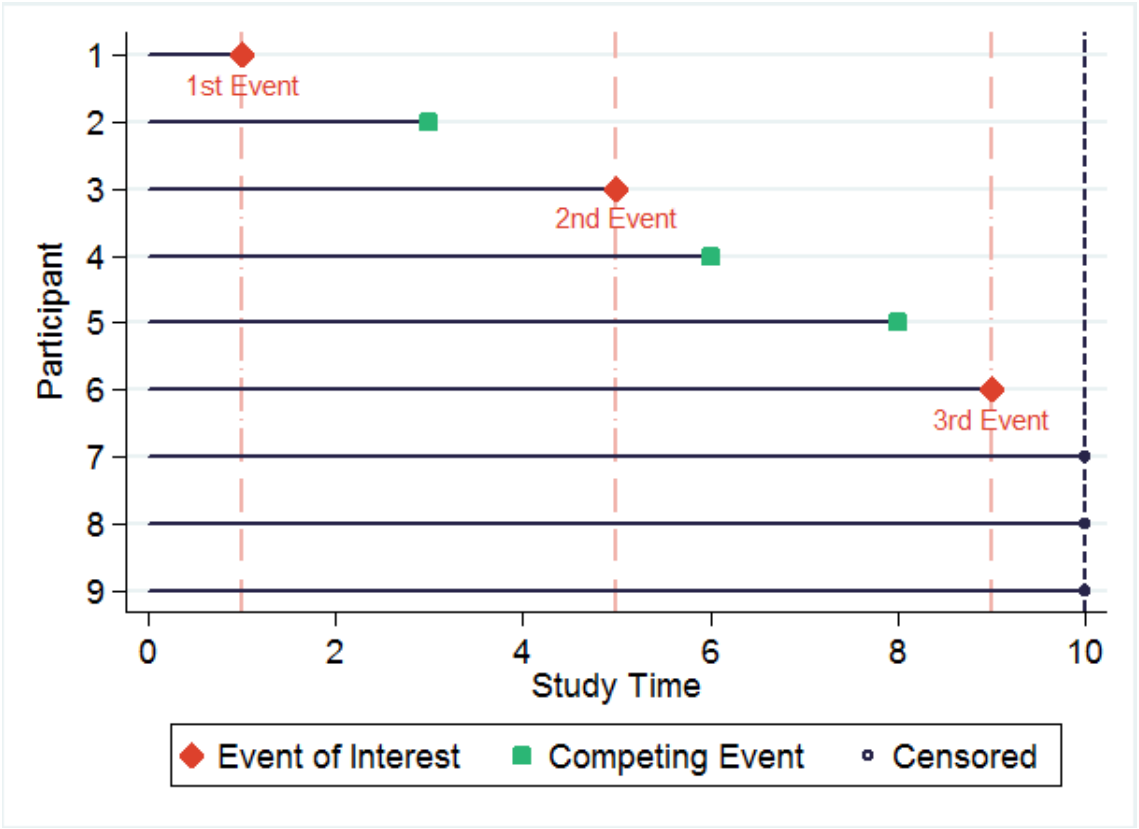
Function name	Notation	Description
All-cause survival	$S(t)$	Cumulative probability of not experiencing any event before time t
All-cause cumulative incidence	$F(t)$	Cumulative probability of experiencing any event before time t
All-cause hazard	$h(t)$	Instantaneous rate of experiencing any event at time t
Cause-specific cumulative incidence	$F_k(t)$	Cumulative probability of experiencing event k before time t
Cause-specific hazard	$h_k(t)$	Instantaneous rate of experiencing event k at time t, if no events previously
Subdistribution hazard	$\lambda_k(t)$	Instantaneous rate of experiencing event k at time t, if no event k previously

1.4.3 Risk sets for cause-specific and subdistribution hazard functions

A key difference between the cause-specific and subdistribution hazard functions is the set of participants who remain in the initial state just before the time point of interest, referred to as the *risk set*. The risk set changes over time as participants experience events or are censored. A graphical representation of the risk sets for the two competing risks approaches is given in Figure 1.9.

In this example, the risk sets are assessed just before the time of an event of interest. The risk set for the cause-specific hazard includes the participants who have not experienced either the event of interest or competing events, and those who experience the competing events are removed (censored). Whereas, the risk set for the subdistribution hazard includes those who have yet to experience the event of interest, and those who experience the competing event remain in the risk set until the end of the study.

Figure 1.9: Cause-specific and subdistribution risk sets



	1st event $t \rightarrow 1$	**2nd event** $t \rightarrow 5$	**3rd event** $t \rightarrow 9$
Cause-specific risk set			
Subdistribution risk set			

At the beginning of the study, the risk set contains all 9 participants.
Over time participants may experience the event of interest (depicted by a red diamond) or the competing event (depicted by a green square).
Those participants that experience a competing event are excluded from the cause-specific risk set, but remain in the subdistribution risk set.
Participants that have left the risk set are faded out.
The risk sets above represent those at risk just before the time of an event of interest ($\mathbf{t} \rightarrow$).

The interpretation of the competing risks hazard measures can be difficult. Similarly, the risk sets associated with the competing risks approaches are challenging to interpret. Consider the example of disease recurrence following breast cancer surgery, again with death as a competing event. When estimating the cause-specific hazard, participants who die transition out of the initial state into the competing state (Figure 1.7), and are thus removed from the risk set. Unlike standard time-to-event analysis, independent censoring is not assumed, as those who die from other causes do not have the same risk of disease recurrence as those who are alive and remain in the initial state/risk set. When estimating the subdistribution hazard, participants who die remain in the initial state/risk set (Figure 1.8). It is impossible for participants that died to subsequently experience disease recurrence, yet being in the risk set implies that they are still at risk. Though this seems counterintuitive, we can consider those who have experienced the competing event as "cured" from the event of interest. These participants act as placeholders (Koller et al., 2012) by remaining in the initial state with no chance of experiencing disease recurrence, thus preventing all participants from transitioning out of the initial state.

1.4.4 Theoretical relationships between key functions

Similar to standard time-to-event analysis, the hazard, survival, and cumulative incidence functions are related in the competing risks setting. These relationships are now described in detail.

1.4.4.1 Relationships with the all-cause survival function.

As established previously, the all-cause survival, cumulative incidence, and hazard functions are equivalent to the functions derived in the standard setting of time-to-event analysis (if all events are combined into a single state). Thus, the relationships which hold in standard time-to-event analysis remain.

By definition, the all-cause hazard function may be written as the sum of all K cause-specific hazard functions:

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T \leq t + \delta | T \geq t)}{\delta}$$

$$= \lim_{\delta \to 0} \frac{\sum_{k=1}^{K} P(t \leq T \leq t + \delta, event = k | T \geq t)}{\delta} = \sum_{k=1}^{K} h_k(t) \qquad \textbf{\textit{Equation 1.22}}$$

As such, the all-cause survival function may be written as;

$$S(t) = exp\left(-\int_0^t h(u)\, du\right) = exp\left(-\int_0^t \sum_{k=1}^{K} h_k(u)\ du\right) \qquad \textbf{Equation 1.23}$$

Estimates of all K cause-specific hazard functions are required to obtain an estimate of the all-cause survival, and thus the all-cause cumulative incidence functions. These functions are key to estimating the absolute risk of the events occurring, the primary interest of prognostic model research.

1.4.4.2 Relationships with the cause-specific cumulative incidence function.

The cause-specific cumulative incidence function represents the risk of experiencing the event of interest over time, and is the key function needed to derive a prognostic model. This function $F_k(t)$, depends on two values at any time t; the probability that the participant remains free from any event until just before time t, and the instantaneous probability that event k occurs at time t given that no event occurred before that time. The first probability is the all-cause survival function $S(t)$, and the second probability is the cause-specific hazard function $h_k(t)$. Thus, the cause-specific cumulative incidence function can be written:

$$F_k(t) = P(T \leq t, event = k) = \int_0^t S(u) \times h_k(u)\, du \qquad \textbf{\textit{Equation 1.24}}$$

Inputting Equation 1.23 for the all-cause survivor function gives;

$$F_k(t) = \int_0^t h_k(u) \exp\left(-\int_0^u \sum_{j=1}^{K} h_j(v)\ dv\right) du \qquad \textbf{\textit{Equation 1.25}}$$

Estimates of all K cause-specific hazard functions are required to estimate the all-cause survival function, to then allow for the estimation of the cause-specific cumulative incidence function. A consequence of this property is that the one-to-one relationship between the rate (cause-specific hazard) and the risk (cause-specific cumulative incidence) of the event of interest is lost; the cumulative incidence of the event of interest depends on the rate of all possible events. Hence, these hazards cannot be interpreted in the same way as those estimated using the standard time-to-event analyses, as they do not directly correspond with the "real world" risk of the event of interest.

To avoid the added complication of the loss of the relationship between the cause-specific hazard and the cause-specific cumulative incidence function, Fine and Gray (Fine and Gray, 1999) introduced the subdistribution hazard function $\lambda_k(t)$. The subdistribution hazard function for event k is equal to the rate of change of the cause-specific cumulative incidence function; derived by differentiating the natural log of the cause-specific cumulative incidence function:

$$\lambda_k(t) = -\frac{d}{dt} \ln[1 - F_k(t)] \qquad \textbf{\textit{Equation 1.26}}$$

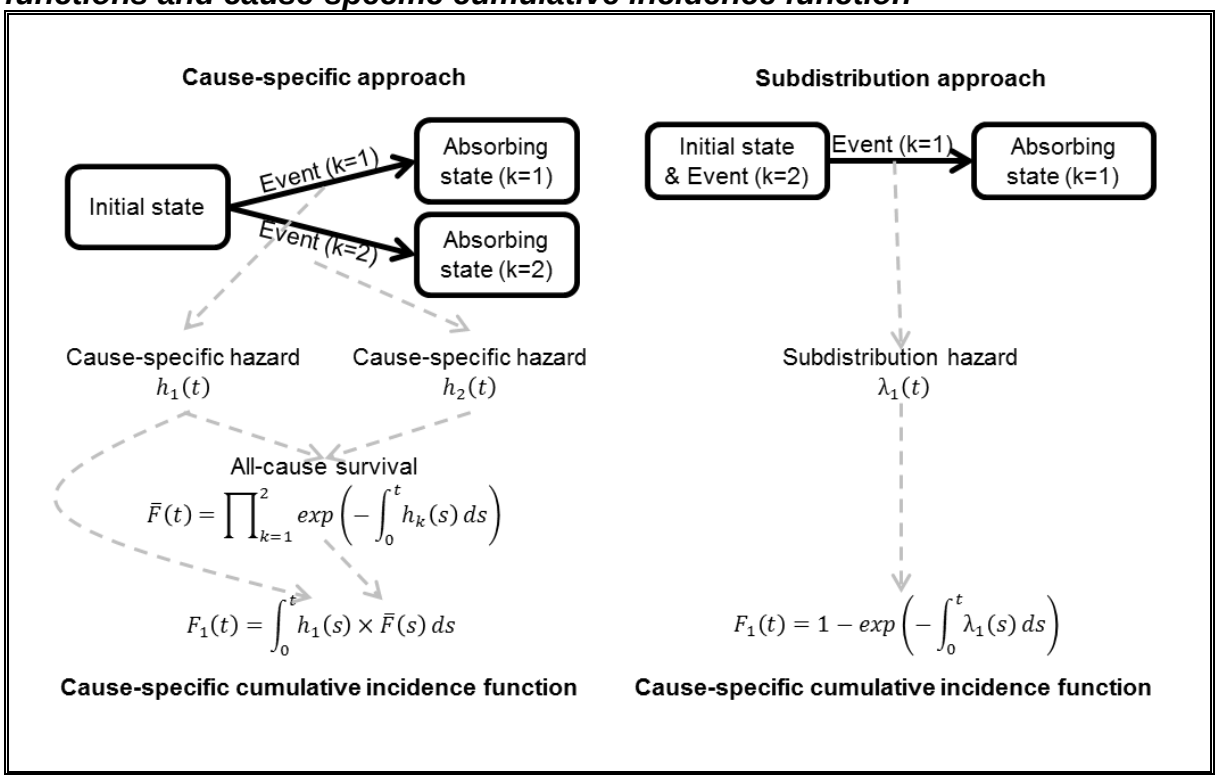
The subdistribution hazard function preserves a one-to-one relationship with the cause-specific cumulative incidence function:

$$F_k(t) = 1 - \exp\left(-\int_0^t \lambda_k(u)\, du\right) \qquad \textbf{\textit{Equation 1.27}}$$

Estimation of the cause-specific cumulative incidence function only requires an estimate of the subdistribution hazard for the event of interest. Hence, it is more straightforward to estimate the cause-specific cumulative incidence function using the subdistribution hazard.

The relationships between the cause-specific and subdistribution hazard functions and the cause-specific cumulative incidence function with two events (the event of interest and the competing event) are depicted in Figure 1.10.

Figure 1.10: Relationships between cause-specific and subdistribution hazard functions and cause-specific cumulative incidence function



From the figure, both the straightforward relationship between the subdistribution function, and the more complex relationship between the cause-specific hazard function, and the cause-specific cumulative incidence function are apparent.

1.4.5 Competing risks regression

The statistical advantages of flexible parametric models over semi-parametric approaches (Snell, 2015) were outlined in Section 1.3.4. Recall, Royston-Parmar flexible parametric models (introduced in Section 1.3.4) estimate a smooth baseline hazard function using restricted cubic splines alongside estimates of regression coefficients, on the log cumulative hazard scale (Royston and Parmar, 2002). These methods have been adapted to incorporate competing events using both the cause-specific (Hinchliffe and Lambert, 2013b) and subdistribution hazards (Lambert et al.,

2017) approaches. The two approaches are discussed below. Again, for simplicity each participant is at risk of experiencing $K = 2$ mutually exclusive events, the event of interest $k = 1$, and a competing event $k = 2$. For regression purposes, we include a vector of prognostic factors $\mathbf{X}_{k,i} = (x_{k,1}, x_{k,2}, \ldots)^T$, for each cause k and each individual i. The model parameters can be estimated using standard maximum likelihood techniques.

1.4.5.1 Cause-specific hazards approach

The flexible parametric modelling approach using cause-specific hazards to estimate cause-specific cumulative incidence functions (Hinchliffe and Lambert, 2013b), models on the log cumulative cause-specific hazards scale. Assuming proportional hazards, the log cumulative cause-specific hazard function for event k with a vector of prognostic factors $\mathbf{X}_{k,i}$ can be written as:

$$\ln\left[H_k\left(t|\mathbf{X}_{k,i}\right)\right] = \text{Spline}\{\ln[t]|\boldsymbol{\gamma}_k, \mathbf{N}_k\} + \boldsymbol{\beta}_k{}^T\mathbf{X}_{k,i} \qquad \textit{Equation 1.28}$$

Where $\text{Spline}\{\ln[t]|\boldsymbol{\gamma}_k, \mathbf{N}_k\}$ represents a restricted cubic spline function (as given in Equation 1.14) with a vector of knot locations $\mathbf{N}_k$ modelling the log cumulative baseline cause-specific hazard function for event k. The resulting vector of regression coefficients $\boldsymbol{\beta}_k$, which corresponds to the vector of prognostic factors $\mathbf{X}_{k,i}$, are estimated using maximum partial likelihood techniques (Hinchliffe and Lambert, 2013b). The regression coefficients can be interpreted as log cause-specific hazard ratios under the proportional hazards assumption, and thus describe the (adjusted) effect of each prognostic factor on the risk of the event of interest when competing events cannot occur.

The cause-specific hazard function is obtained by differentiating the cumulative cause-specific hazard function with respect to time, and thus involves the derivatives of the restricted cubic spline functions:

$$h_k(t|\mathbf{X}_{k,i}) = \frac{d}{dt}\exp\left(\text{Spline}\{\ln[t]|\boldsymbol{\gamma}_k, \mathbf{N}_k\} + \boldsymbol{\beta}_k{}^T\mathbf{X}_{k,i}\right) \quad \textit{Equation 1.29}$$

Estimation of each cause-specific hazard function can be obtained by fitting separate models for each of the K events, while censoring all other events. Alternatively, it is possible to simultaneously estimate all K cause-specific hazard functions using a multi-state data format. Though the two approaches give the same results, simultaneous estimation is considered more flexible. It allows for shared coefficient estimates and baseline hazard functions across events, and allows for convenient testing and comparison of coefficient estimates across event types (Lunn and McNeil, 1995).

To model all K cause-specific hazard functions simultaneously, the dataset to which the model is applied must be expanded to mimic the K different datasets used if the analyses were performed separately. This is best illustrated using an example. When investigating the risk of disease recurrence following breast cancer surgery, death is a competing event, as such there are $K = 2$ mutually exclusive events. The original data may look like that depicted in Figure 1.11**a**, with participant 1 experiencing disease recurrence at 3.1 years, participant 2 dying at 4.1 years, and participant 3 being censored at 5.9 years. To simultaneously analyse all events, the dataset is expanded to the multi-state data, depicted in Figure 1.11**b**. Now each participant has $K = 2$ rows, one for each of the mutually exclusive events. Participant 1 experiences disease recurrence at 3.1 years (row 1), but is censored at 3.1 years for death (row 2). Participant 2 is censored for disease recurrence at 4.1 years (row 3), but dies at 4.1 years (row 4). Finally, participant 3 is censored at 5.9 years for both of the events (rows 5 and 6).

Figure 1.11: Original and multi-state datasets for competing risks analysis

a) Original data (wide format)

ID	Age	Time	Disease Recurrence	Death
1	34	3.1	1	0
2	56	4.1	0	1
3	42	5.9	0	0
…	…	…	…	…

b) Multi-state data (long format)

Row	ID	Age	Time	Event	Status	Row
1	1	34	3.1	Disease Recurrence	1	1
2	1	34	3.1	Death	0	2
3	2	56	4.1	Disease Recurrence	0	3
4	2	56	4.1	Death	1	4
5	3	42	5.9	Disease Recurrence	0	5
6	3	42	5.9	Death	0	6
	…	…	…	…	…	

In prognostic model research, the key measure of interest is the cause-specific cumulative incidence function, as this returns individual estimates of absolute risk predictions. As discussed previously, estimating the cause-specific cumulative incidence function using the cause-specific approach requires estimates of all K cause-specific hazard functions. To estimate for the cause-specific cumulative incidence function for event k with a vector of prognostic factors $\mathbf{X}_{k,i}$ recall Equation 1.25:

$$\widehat{F_k}(t|\mathbf{X}_{k,i}) = \int_0^t \widehat{h_k}(u|\mathbf{X}_{k,i}) \exp\left(-\int_0^u \sum_{j=1}^K \widehat{h_j}(v|\mathbf{X}_{j,i})\ dv\right) du \qquad \textit{Equation 1.30}$$

The integral in the above equation cannot be solved analytically, thus additional methods, such as numerical integration (Hinchliffe and Lambert, 2013b) or the simulation approach (Fiocco et al., 2008, Crowther and Lambert, 2017), are required. Briefly, the simulation approach simulates a large sample of participants, and calculates a transition probability matrix using Nelson-Aalen estimators of the cumulative cause-specific hazard function $H_k(t)$. The simulated participants iterate through the transition probability matrix until they all either experience an event or are censored (at a specified maximum follow-up time). Cause-specific cumulative

incidence function estimates for event k given a prognostic factor vector $\mathbf{X}_{k,i}$ are calculated as the proportion of simulated participants with the same vector of prognostic factor values that experience event k. The cause-specific flexible parametric model can be fitted to all causes simultaneously using the expanded multi-state dataset (Crowther and Lambert, 2017). An example of a published model developed using the cause-specific approach is provided in Box 1.4.

1.4.5.2 Subdistribution hazards approach

The flexible parametric modelling approach using subdistribution hazards to estimate cause-specific cumulative incidence functions (Lambert et al., 2017), models on the log cumulative subdistribution hazards scale. The cumulative subdistribution hazard function is defined as $\Lambda_k(t) = \int_0^t \lambda_k(u)du$ for continuous time distributions. Assuming proportional subdistribution hazards[i], the log cumulative subdistribution hazard function for event k with a vector of prognostic factors $\mathbf{X}_{k,i}$ is written:

$$\ln\left[\Lambda_k\left(t|\mathbf{X}_{k,i}\right)\right] = \text{Spline}\{\ln[t]|\boldsymbol{\gamma}_k, \mathbf{N}_k\} + \boldsymbol{\beta}_k{}^T\mathbf{X}_{k,i} \qquad \textit{Equation 1.31}$$

In this case the restricted cubic spline function is modelled on the log cumulative baseline subdistribution hazard scale for event k. The resulting vector of regression coefficients $\boldsymbol{\beta}_k$, which corresponds to the vector of prognostic factors $\mathbf{X}_{k,i}$, are estimated by maximising the weighted partial likelihood function (Lambert et al., 2017). The regression coefficients can be interpreted as log subdistribution hazard ratios under the proportional subdistribution hazards assumption. Thus, describe the effect of each prognostic factor on the risk of the event of interest, adjusted for the occurrence of competing events.

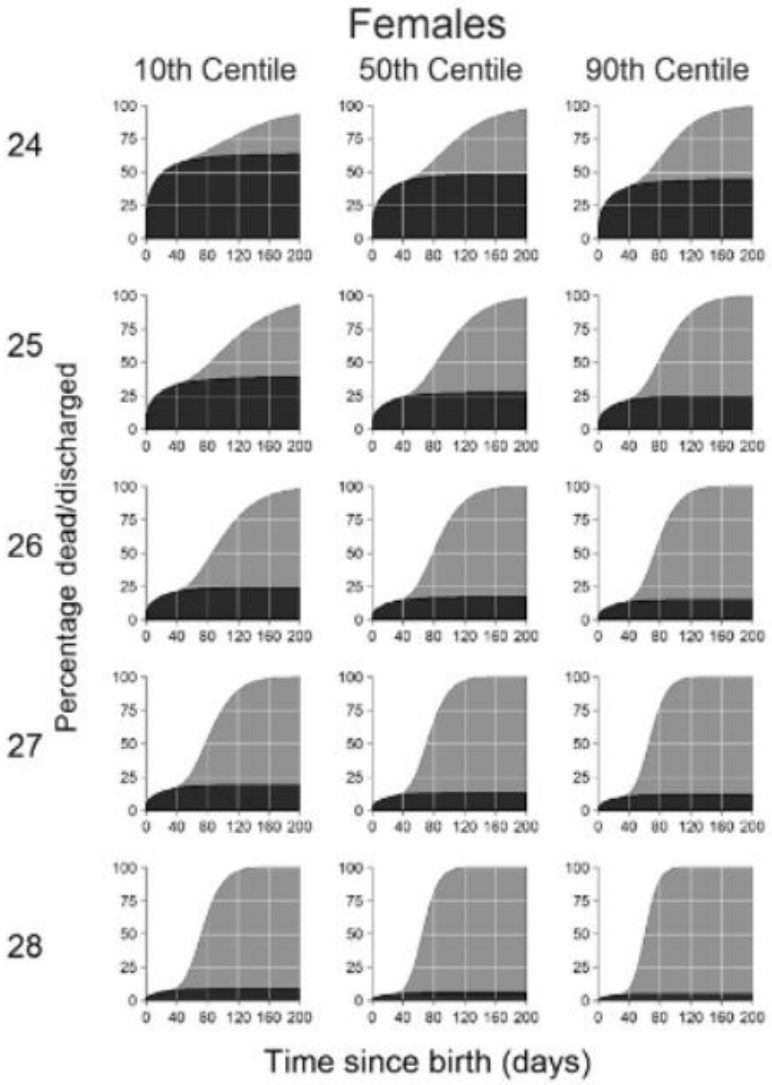
[i] Note if the proportional cause-specific hazards assumption holds then the subdistribution hazards cannot be proportional over time (Beyersmann et al., 2009, Grambauer et al., 2010)

Box 1.4: Example of a model developed using cause-specific approach

The time to death or discharge in neonatal care was examined in a recent study (Hinchliffe et al., 2013) using flexible parametric modelling and the cause-specific approach. The model was developed using retrospective data from 2,723 babies born at 24-28 weeks gestational age, admitted to neonatal care. Flexible parametric methods were used to analyse death and discharge alive as competing events.

Gestational age (weeks), sex, and birthweight (centiles) were found to significantly effect the time to death or discharge. Cause-specific hazard ratios were not reported.

The cause-specific cumulative incidence for death and discarge alive for female babies admitted to neonatal care are provided below for different gestational ages (top to bottom) and birthweights (left to right).



Absolute probabilities for death (black) and discharge (grey) for female babies admitted to neonatal care, by gestational age and birthweight centile. (Hinchliffe et al., 2013)[i]

[i] Reproduced from "Modelling time to death or discharge in neonatal care: an application of competing risks" Hinchliffe et al., 426-433, 27, 2013, with permission from John Wiley and Sons.

Estimation of the subdistribution hazard function for the event of interest $k = 1$ can be performed without needing to model the competing events. This is achieved by utilising the subdistribution risk set, in which participants who experience competing events remain in the risk set but are unable to experience the event of interest (Figure 1.9). If proportional subdistribution hazards are not biologically plausible, alternative link functions may be used (e.g. a logit link gives a proportional odds model) (Lambert et al., 2017), or time dependent effects can be modelled by incorporating interactions between the prognostic factors and the restricted cubic spline function (Hinchliffe and Lambert, 2013b).

The cause-specific cumulative incidence is of key interest in prognostic model research. This can be estimated using the subdistribution approach by incorporating the subdistribution hazard function (above) into Equation 1.27:

$$\widehat{F_k}(t|\mathbf{X}_{k,i}) = 1 - \exp\{-\exp(\text{Spline}\{\ln[t]|\boldsymbol{\gamma}_k, \mathbf{N}_k\} + \boldsymbol{\beta}_k{}^T\mathbf{X}_{k,i})\} \qquad \textbf{\textit{Equation 1.32}}$$

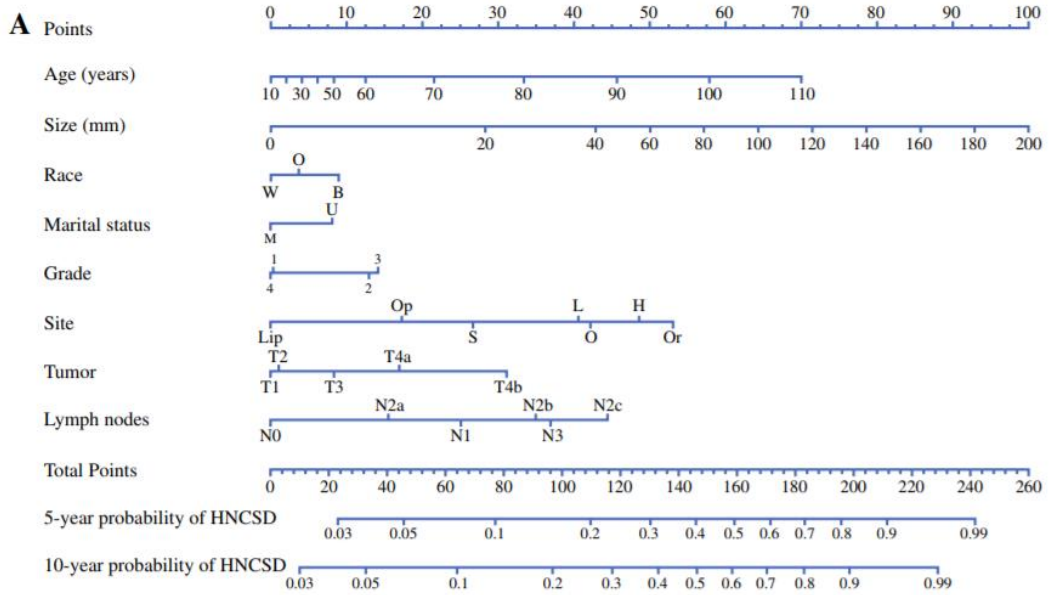
While it is not necessary to model any event other than the event of interest, it is possible to fit the subdistribution flexible parametric model to each cause, either separately or simultaneously (Lambert et al., 2017). An example of a published prognostic model developed using the subdistribution approach is provided in Box 1.5.

Box 1.5: Example of a prognostic model developed using the subdistribution approach

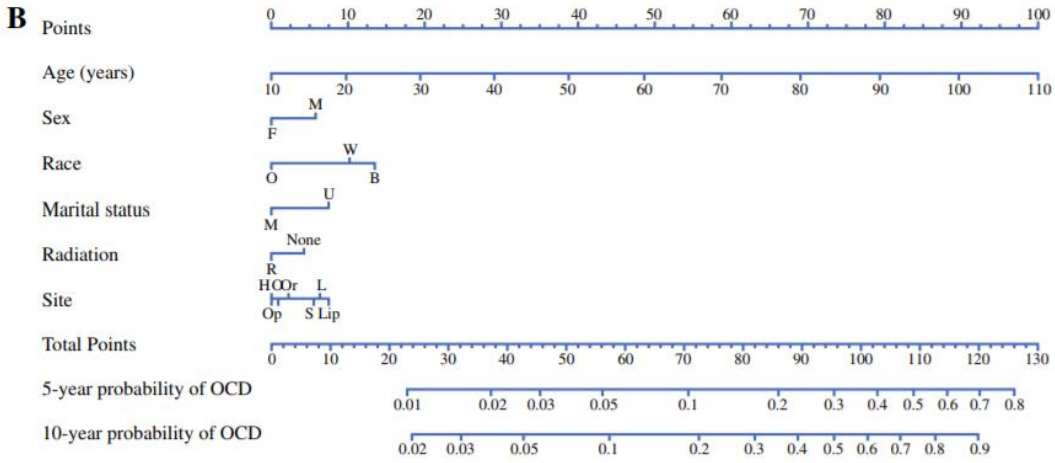
The time to cancer-specific mortality, with other causes death as a competing event, in patients with head and neck squamous cell carcinoma was examined in a recent study (Shen et al., 2015) using flexible parametric modelling and the subdistribution apparoach. The model was developed using a cohort of 23,494 patients with head and neck squarmous cell carcinoma.

The following prognostic factors were investigated for each cause: age (years), race (white, black, other), marital status (unmarried, married), radiotherapy (none, yes),.tumour size (mm), grade, T and N classifications, and site (lip, oral cavity, salivary gland, oropharynx, hypopharynx, larynx, other).

Cause-specific cumulative incidences were estimated seperately for each cause. Nomograms for predicting 5 and 10 year probabilities of each cause of death are below:



Nomogram for predicting 5 and 10 year probability of cancer-specific death. (Shen et al., 2015)[i]



Nomogram for predicting 5 and 10 year probability of other causes death. (Shen et al., 2015)

[i]Adapted by permission from Springer Nature: Annals of Surgical Oncology "Cancer-Specific Mortality and Competing Mortality in Patients with Head and Neck Squamous Cell Carcinoma: A Competing Risk Analysis", Shen et al., (2015)

1.4.5.3 Differences between cause-specific and subdistribution approaches

A summary of the key differences between the cause-specific and subdistribution modelling approaches is provided in Table 1.2.

Table 1.2: Differences between cause-specific and subdistribution modelling approaches

	Cause-specific approach	Subdistribution approach
Model assumptions	Proportional cause-specific hazards $h_k(t)$.	Proportional subdistribution hazards $\lambda_k(t)$.
Risk sets	The same as standard time to event analysis: contains participants who have not experienced either the event of interest or any competing event. Participants who experience competing events are censored.	Differs from standard time-to-event analysis: contains participants who have not experienced the event of interest. Participants who experience competing events remain in the risk set, and are "cured" from the event of interest.
Interpretation of hazard (prognostic factor associations)	The cause-specific hazard measures the direct association of prognostic factors on the event of interest, assuming the competing events cannot occur (i.e ignoring the indirect effects of the competing events).	The subdistribution hazard measures the association of the prognostic factors on the "real world" risk of the event of interest, incorporating the indirect effects of the competing event.
Link to cumulative incidence (absolute risks)	There is no 1:1 relationship between the cause-specific hazard and the cumulative incidence function. Estimation of ALL cause-specific hazard functions is required to obtain absolute risk estimates.	There is a 1:1 relationship between the subdistribution hazard and the cumulative incidence function. Need only estimate the subdistribution hazard function for the event of interest to obtain absolute risk estimates.
When should the approach be used?	Prognostic factor research	Prognostic model research
Main advantage	Measures direct associations so aids understanding of eitiological questions.	Straightforward link to cumulative incidence function, so makes it easier to model just the event of interest.

1.5 Prognostic model development

This thesis applies the described time-to-event analysis methods to develop prognostic models in the competing risks setting. Now that the framework for standard and competing risks time-to-event analysis have been described, it is important to provide a general overview of other key issues that arise when developing prognostic models. Statistical considerations which are generally advocated during the development of prognostic models include: evaluating data quality; manipulation and selection of candidate prognostic factors; testing model assumptions; and managing overfitting and optimism (internal validation) (Moons et al., 2012b, Royston et al., 2009, Steyerberg, 2008, Steyerberg et al., 2013). An overview of each of these is now given.

1.5.1 Evaluating data quality

The quality of the data is evaluated to ensure the data is fit for purpose with minimal measurement error (Royston et al. 2009), to enhance the reliability of the prognostic model (Steyerberg et al., 2013). Data should ideally come from a prospective cohort study with standardised measurement of the outcome and candidate prognostic factors (Moons et al., 2012b).

Missing values are common in prognostic model research (Moons et al., 2012b), and numerous statistical techniques exist to manage these. Complete case analysis excludes any participant with missing information, resulting in a loss of statistical power and often invalid parameter estimates (Royston et al., 2009, Moons et al., 2012b), thus is only advised when few observations (say <5%) are missing (Harrell, 2015). Imputation methods utilise multivariable models containing both dependent and independent variables to impute plausible values to replace missing information, these are typically more efficient than complete case analysis (White and Royston, 2009). Multiple imputation methods repeat imputations a number of times to produce multiple completed datasets (Harrell, 2015), which incorporates the uncertainty from the

imputed values into the results. Each imputed dataset is analysed using standard methods, and analysis results are combined using Rubin's rules (Rubin, 2004); Those applicable to prognostic model research (Marshall et al., 2009) are outlined in Table 1.3. Multiple imputation using chained equations (Buuren and Oudshoorn, 2000) utilises iterative applications of regression imputation to recover missing values under a missing at random assumption.

Table 1.3: Rubin's rules for combining estimates from multiply imputed data

Parameter of Interest		Equation for combining estimate
Regression coefficient	$\overline{\beta}$	$\overline{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_i$
Estimated within imputation variance	$\overline{U}$	$\overline{U} = \frac{1}{M}\sum_{i=1}^{M}U_i$
Between imputation variance	B	$B = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\beta}_i - \overline{\beta}\right)^2$
Total variance	T	$T = \overline{U} + \left(1 + \frac{1}{M}\right)B$
Wald test statistic	W	$W = \frac{\overline{\beta}^2}{T}$
Degrees of Freedom	v	$v = (M-1)\left(1 + \frac{\overline{U}}{\left(1+\frac{1}{M}\right)B}\right)^2$
P-value	p	$p = F_{1,v}(W)$ F-Distribution test.

The above rules, known as Rubin's rules (Rubin, 2004), are given for combining estimates in $M > 1$ imputed datasets. Where $\hat{\beta}_i$ represents the estimated regression coefficient in the i^{th} imputed dataset, and U_i represents its variance.
The application of these rules specifically for prognostic model research is discussed further in (Marshall et al., 2009).

It is advised when imputing missing information with a time-to-event outcome that outcome information (a binary variable indicating whether the event occurred and an estimate of the cumulative hazard function) is incorporated into the chained equations (White and Royston, 2009). When competing events are present the chained

equations should include a categorical event indicator[i], and Nelson-Aalen estimators for each of the K cumulative cause-specific hazard functions over time (Bartlett and Taylor, 2016). Logically, when using the cause-specific approach information on all K events are incorporated into the chained equations, as all K events are modelled to estimate the cause-specific cumulative incidence function. As the subdistribution approach only requires the event of interest to be modelled, the chained equations need only include a binary event indicator for the event of interest, and Nelson-Aalen estimate of the cumulative subdistribution hazard function for the event of interest.

1.5.2 Manipulation and selection of candidate prognostic factors

The term “candidate prognostic factor” refers to any prognostic factor which is considered during the development of the prognostic model (Moons et al., 2012b). Existing prognostic factors with known predictive ability, evidenced by prognostic factor research, are usually considered as candidates in prognostic model development (Royston et al., 2009). Manipulation of candidate prognostic factors may be required to aid statistical modelling. For instance, new prognostic factors can be created through the combination of existing ones, such as BMI created from height and weight. Categories within ordered categorical prognostic factors that contain small numbers of participants may be collapsed to give more stable results (Steyerberg 2008). Continuous prognostic factors may require manipulation if not thought to have a linear relationship with the outcome. Transformation of the prognostic factor (for example using fractional polynomial functions) is preferred over categorisation, as more predictive information is retained (Moons et al., 2012b)(Royston et al. 2009). For prognostic models with time-to-event outcomes, centering of continuous prognostic factors helps to ensure a meaningful baseline hazard function (Royston and Altman, 2013) aiding interpretation of relative risks. The reference category in categorical

[i] A categorical variable which indicates which event occurred during the study period; equal to 0 for no event (censored), 1 for the event of interest, 2,…,K for all other competing events.

factors should also be selected with care (Steyerberg, 2008) to ensure a representative baseline to which relative risk comparisons are made (the importance of which is highlighted in Box 1.4).

Often more candidate prognostic factors are available than can sensibly be used in a prognostic model (Royston et al., 2009), and selection methods are required to reduce the quantity. Parsimonious models, which reach a suitable level of prediction but contain fewer prognostic factors, are preferred. Strongly correlated prognostic factors can be removed from models, as they contribute little independent information and explain the same variation (Harrell, 2015, Royston et al., 2009). Numerous selection methods are used in practice (Heinze et al., 2018), however there is no consensus on which is the “best” approach (Royston et al., 2009). Fitting the full model (i.e. including all candidate prognostic factors) avoids selection bias, reduces the potential for overfitting, and leads to reliable confidence intervals, though perhaps at the expense of an unnecessarily complex model (Royston et al., 2009). Automated selection algorithms are commonly applied to reduce the number of candidate prognostic factors included in the final multivariable model, of those available the backward selection method is recommended (Collins et al., 2015, Moons et al., 2015). Backwards elimination is one such algorithm, in which initially a full model is fitted. An iterative process, in which the least statistically significant prognostic factor is eliminated and the model is re-fitted, is repeated until all remaining prognostic factors are significant at a pre-specified significance level. Alternative approaches which combine variable selection with adjustment for overfitting, such as the LASSO, LARS and elastic net methods, may be useful for variable selection, but are not considered in this thesis.

When competing events are present, the selection of prognostic factors may also depend on the analysis approach. Firstly, all K events are modelled for the cause-specific approach; hence it is possible for each of the cause-specific hazard model to

contain different combinations of prognostic factors. Secondly, the differences in the subdistribution and cause-specific hazards could result in different sets of prognostic factors being selected for the final model, depending on the approach taken.

To ensure clinical credibility, prognostic models should include clinically important prognostic factors (i.e. those considered to be prognostic by clinicians) regardless of statistical significance (Wyatt and Altman, 1995). Additionally, prognostic factors representing treatments or therapies should ideally be included in the final prognostic models, to circumvent poor prognostic performance if applied to future patients not receiving those treatments (Groenwold et al., 2016).

1.5.3 Testing model assumptions

General assumptions made in multivariable regression modelling equations include linearity and additivity. Linearity refers to the assumption that the effect of a single unit increase in a continuous prognostic factor on the log cumulative hazard function is constant. If this assumption is not appropriate, non-linear relationships may be incorporated by transforming the prognostic factors using fractional polynomial functions or splines (Harrell, 2015). Additivity refers to the assumption that the effects of one prognostic factor are independent of another. If this assumption is not appropriate, interaction terms between prognostic factors can be incorporated into the model (Steyerberg 2008). Additionally, prognostic models with time-to-event outcomes assume both proportional hazards and non-informative censoring.

1.5.4 Managing overfitting and optimism

Statistical overfitting is present when a model is too complex for the amount of information in the data (Harrell, 2015), such that the model is too closely adapted to the data in which it was developed and regression coefficients are overestimated (Royston et al., 2009). Optimism in prognostic models is less likely if developed using a sufficient sample size in which few candidate prognostic factors are tested relative

to the number of events (Riley et al., 2019b, Royston et al., 2009, Steyerberg et al., 2013). In particular, it is often recommended that there should be at least ten events per candidate prognostic factor (Peduzzi et al., 1995, Peduzzi et al., 1996). Optimism is identified through internal validation techniques, discussed below, and can be adjusted for using shrinkage methods for a more robust model for the intended population (Steyerberg, 2008).

1.6 Validation of prognostic models

In prognostic model research, the term validation refers to the statistical process for evaluating the predictive performance of a developed prognostic model (Altman and Royston, 2000). It is widely accepted that a prognostic model should not be used in clinical practice without evidence that it performs a useful role (Moons et al., 2009a), i.e. has good validation performance. The reproducibility of the prognostic model and potential optimism in model performance estimates are evaluated in the same sample of participants used to develop the model, known as *internal validation*. Whereas, the robustness and generalisability (also called transportability) of the model are evaluated in an independent sample of plausibly related participants, known as *external validation* (Steyerberg, 2008, Royston and Altman, 2013). Internal validation is recommended as a pre-requisite for prediction model development, particularly in the case of limited data (Collins et al., 2015). Validation statistics are utilised to evaluate the predictive performance of the model by assessing prediction accuracy (calibration) and ability to reliably distinguish between those who do and do not experience an event (discrimination) in practice. Methods for the validation of prognostic models with time-to-event outcomes are *"not particularly well worked out in the literature"* (Royston and Altman, 2013) due to the additional challenge of incorporating time and censoring information into the model evaluation. Often, prognostic models are validated for predictions at a certain time point rather than over the entire prediction horizon.

1.6.1 Internal validation

Internal validation studies utilise all, or subsets of, the participants in which the prognostic model was developed to evaluate model reproducibility and optimism. The apparent performance of a prognostic model can be evaluated using the exact sample of participants from which the model was derived (Steyerberg, 2008). However, as model parameters have been optimised to fit this exact sample, the apparent predictive performance measures are likely to be optimistic; internal validation is used to assess the likely optimism in the model performance and to correct for the resulting overfitting of the model (Royston & Altman 2013). This is commonly achieved by splitting the sample of participants into development and validation sets (Altman et al., 2009). Random splits are generally not advised, as the randomness leads to homogeneity between the sets resulting in optimistic results (Altman et al., 2009). Non-random splitting may be preferable as it allows for non-random variation between the datasets (Collins et al., 2015), however splitting also leads to a loss of power during model development (Moons et al., 2012b). Resampling techniques, such as cross-validation and bootstrapping, evaluate the predictive performance of a model after development (Collins et al., 2015), allowing the model to be developed using the full participant sample (Altman et al., 2009).

Bootstrapping mimics the study sampling process by drawing, with replacement, from the sample of participants in the study to create a new sample containing the same number of participants as the original. It is useful for validating the modelling process used to develop the original prognostic model, as the modelling process is repeated in each bootstrap sample. For internal validation, a bootstrap sample is drawn and a model is developed using the same modelling process as used to develop the original model. The predictive performance of this model is assessed in the bootstrap sample (bootstrap apparent performance) and the original sample (test performance). The difference of the bootstrap apparent and test performance measures indicates the

optimism from the original model (Moons et al. 2012). This process is repeated using a number of bootstrap samples and an average optimism estimate for each performance measure is calculated. The predictive performance of the original model can be adjusted for overfitting and optimism by subtracting the average optimism estimates from the original model's apparent performance measures (Moons et al., 2012b). Additionally the original model may be adjusted for optimism by multiplying the estimated regression coefficients (for the predictors) by a uniform shrinkage factor, equivalent to the mean optimism adjusted calibration slope, to produce an optimism adjusted model (Moons et al., 2012b, Steyerberg, 2008). Following optimism adjustment, the baseline hazard function should be re-estimated for prognostic models with time-to-event outcomes, to retain overall calibration.

1.6.2 External validation

While internal validation can highlight possible fragilities such as overfitting and optimisation, internal data do not provide the heterogeneity expected during external applications of the model, for this external validation is required (Royston, 2010). External validation refers to the evaluation of the predictive performance of a model in an independent sample of plausibly related participants, and is essential to assess the model's clinical value (Altman et al., 2009, Collins et al., 2016). External validation may be used to demonstrate the appropriateness of application of a prognostic model across different clinical settings, populations, and subgroups of interest, or can alternatively identify where model updating and recalibration strategies are required to improve predictive performance (Riley et al., 2016, Altman et al., 2009). Though an external validation study should be independent, it should also include the same set of prognostic factors, comparably defined outcomes with similar follow-up times, and the same baseline health state as the development study (Royston, 2010). External validation is often a more rigorous assessment of predictive performance, especially

in terms of generalisability, compared to internal validation, and often worse predictive performance is observed than in the original model development study.

1.6.3 Validation statistics

Various statistics are used to quantify the predictive performance of prognostic models. Key validation statistics for prognostic model research assess how accurately the model predicts the observed risk of participants (calibration) and how well the model distinguished between those who do and do not experience the event of interest (discrimination) (Steyerberg, 2008). Other measures of predictive performance exist; however, a comprehensive review of these methods is beyond the scope of this thesis. Instead only those which are relevant to, and thus utilised in, this thesis are discussed below.

1.6.3.1 Measures of calibration

Calibration reflects the prediction accuracy of the model (Royston and Altman, 2013), it measures how accurately the expected risks from a prognostic model predict the observed risks of the participants. Calibration of prognostic models with time-to-event outcomes is only possible if the models contain an estimate of the baseline survival function, as this is required to calculate participants' absolute risks (Royston and Altman, 2013). For time-to-event outcomes the expected and observed risks of the event of interest are evaluated both over time and within specified time periods (Moons et al., 2012b). Calibration statistics which can be utilised for models with time-to-event outcomes include: overall calibration; the expected/observed ratio (at particular time-points); and the calibration slope.

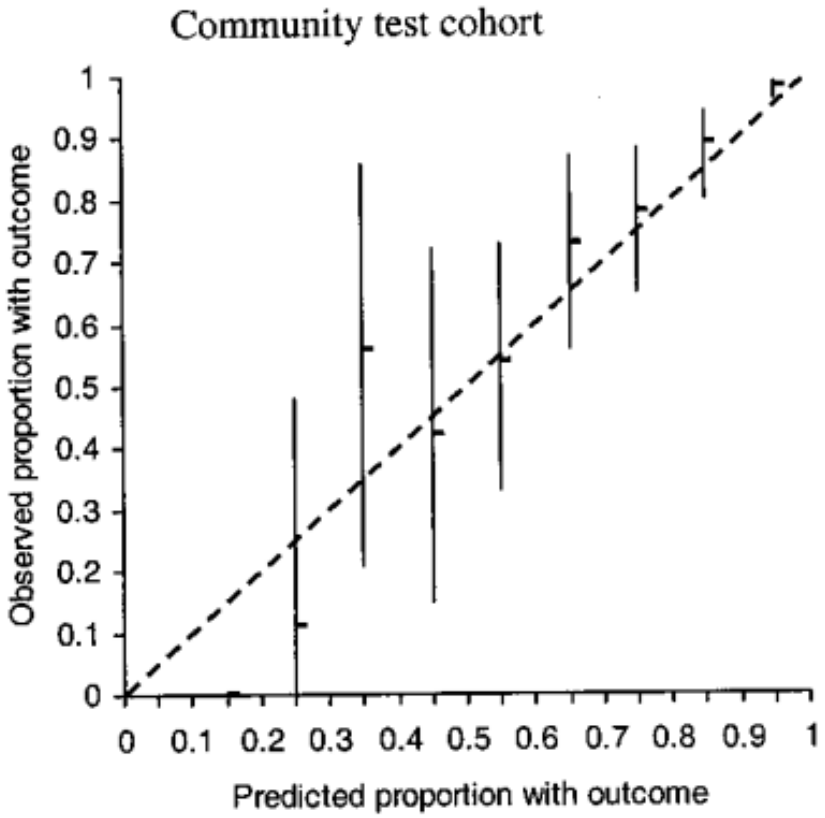
Overall calibration (also known as calibration-in-the-large) compares the expected probability of events predicted by the prognostic model to the observed probability of events in the study, over the study time period. As censoring is likely to be present in prognostic model studies with time-to-event outcomes, estimates of

expected and observed risks should account for censoring. The expected risk is represented by the predicted cumulative incidence curve from the fitted model, and the observed risk is represented by the 1– Kaplan-Meier curve of the study participants' cumulative risk. Overall calibration is assessed by overlaying the curves on the same plot (Royston and Altman, 2013), if the curves are similar the model is "well calibrated". This can be assessed for the sample as a whole (calibration-in-the-large) or for risk groups, created by grouping participants with similar predicted risks, to assess calibration at different levels of risk. Often risk groups are created by splitting at deciles of predicted probabilities (Steyerberg, 2008), however it is recommended that risk groups contain a minimum of 50 participants for stable Kaplan-Meier estimation (Harrell, 2015). The expected/observed ratio (E/O) assesses overall calibration at specified time points. Again E/O may be reported for the entire validation sample or over risk groups. The ratio of the expected and observed probabilities of an event should be close to one for a well calibrated model.

Calibration plots graphically depict the expected and observed probabilities of an event occurring prior to a specified time point for risk groups (Altman et al., 2009), as depicted in Box 1.6. The slope of a fitted line in a calibration plot is referred to as the calibration slope (Steyerberg, 2008). When assessing prognostic models with time-to-event outcomes the calibration slope can be assessed as an average slope over time by estimating the regression coefficient of a model containing the linear predictor from the prognostic model as the only variable (Royston and Altman, 2013). A well calibrated model produces a calibration slope estimate close to one and E/O close to one at most time-points. Methods of recalibration may be utilised to improve prognostic models with poor calibration to improve model fit (Royston, 2010).

When competing events are present, observed risks are calculated using cumulative incidence estimates (rather than survival estimates), as these account for censoring and the competing events (Wolbers et al., 2009).

Box 1.6: Example of a calibration plot



Calibration plot from the external validation of a prognostic model for 30 day risk of mortality following stroke developed using Cox regression (Counsell et al., 2002)[i].
Deciles of predicted probability were used to form groups in calibration plot.
The diagonal dashed line represents perfect calibration and the vertical lines represent 95% confidence intervals for observed risk in each group.

1.6.3.2 Measures of discrimination

Discrimination refers to the extent to which predicted risk estimates distinguish between different patient prognoses (Royston and Altman, 2013). For prognostic models with time-to-event outcomes, discrimination not only distinguishes between those who do and do not go on to experience the event, but also should distinguish between the times at which the events occur. Groups of participants with higher predicted risks should have higher event rates and experience events sooner than those with lower predicted risks (Royston and Altman, 2013). Discrimination statistics

[i] Adapted by permission from Wolters Kluwer Health, Inc. "Predicting Outcome After Acute and Subacute Stroke", Counsell et al., Stroke. 2002; 33:1041-1047

utilised for models with time-to-event outcomes include Harrell's C-index (Harrell et al., 1982), and Royston and Sauerbrei's D-statistic and R_D^2 (Royston and Sauerbrei, 2004).

Harrell's C-index (Harrell et al., 1982) is the probability that, of two randomly chosen participants, the one with the highest expected risk will experience the event first. Not all pairs of participants are evaluable; if neither of the participants experience the event during study follow-up it is not possible determine which will experience the event first. Similarly, if a participant is censored before the other experiences the event, it is not possible to determine the order of the pairing. Thus, Harrell's C-index is known to be biased in instances with heavy censoring, (Royston and Altman, 2013).

Royston and Sauerbrei's D-statistic is a measure of prognostic separation (Royston and Sauerbrei, 2004). This can be interpreted as the log hazard ratio comparing two equal-sized groups, created by splitting the sample using the median value of the estimated linear predictor from the prognostic model (Riley et al., 2016). This is achieved by utilising the linear predictor from the prognostic model to calculate each individual's linear predictor value. These values are ordered, and corresponding standard normal order statistics (rankits) are calculated, and then scaled by a factor $\kappa = \sqrt{8/\pi}$. The scaled rankits are then regressed on the outcome, the resulting estimated regression coefficient is the D-statistic (Royston and Sauerbrei, 2004). Higher values of the D-statistic represent more separation, thus greater discrimination. Once calculated the D-statistic can be incorporated into a generalisation of the multiple correlation coefficient R_D^2, representing the proportion of explained variation on the log relative hazard scale (Royston and Altman, 2013) as follows:

$$\boldsymbol{R_D^2 = \frac{D^2/\kappa^2}{\sigma^2 + D^2/\kappa^2}, \quad \sigma^2 = \frac{\pi^2}{6}} \qquad \textbf{\textit{Equation 1.33}}$$

1.7 Aims and overview of the thesis

The aim of this thesis is to improve understanding of the influence of competing risks issues in prognostic model research. In particular, this thesis aims to;

1. Investigate the presence of competing events in prognostic model research studies, specifically in systematic reviews and development studies.
2. Develop, externally validate, and compare prognostic models for the risks of antenatal adverse events in pre-eclampsia pregnancies which do and do not appropriately account for the competing risk of delivery.
3. Evaluate the impact of not correctly accounting for competing events on prognostic performance measures,
4. Use simulation studies to examine the credibility of a proposed rule of thumb for deciding when competing risks should be accounted for.

This thesis is comprised of eight chapters. The first aim is addressed in Chapters 2 & 3, where evaluations of published prediction model literature investigate the presence, reporting, and management of competing events. The second aim is addressed in Chapters 4 to 6, in which semi-parametric and flexible parametric regression methods are utilised to develop, and then externally validate, prognostic models for patients with pre-eclampsia. The third and fourth aims are addressed in Chapter 7, in which multiple simulation studies are utilised to investigate the effect of not appropriately accounting for competing events when evaluating prognostic model performance measures. The thesis ends with a discussion of key findings, here limitations within the thesis and plans for further research are detailed. An outline of the chapters is given below.

In **Chapter 2**, the presence and reporting of competing events, and potential for competing risks bias, in systematic reviews of prediction model studies are investigated. The potential for competing risks bias in the included studies is examined through assessment of the outcome of interest, characteristics of study populations,

and the prediction time spans of included models. The reporting of competing events and use of quality assessment tools within the systematic reviews are also investigated.

Chapter 3 reviews the presence, reporting, and management of competing events in prediction model development studies that are likely to be affected by competing events. Again, the potential for competing risks bias is examined through assessment of the outcome of interest, characteristics of study populations, and prediction time spans of the included models. The reporting and management of competing events is inspected by determining whether statistical regression methods are applied to appropriately account for the competing events, and whether participants who experienced competing events are censored or excluded from the studies.

In **Chapter 4**, both Cox proportional hazards and Royston-Parmar flexible parametric modelling methods are applied to develop and internally validate prognostic models which predict the risk of antenatal adverse events in pregnancies with early-onset pre-eclampsia. Multiple imputation methods are implemented to manage missing data, fractional polynomial associations between the prognostic factors and outcome are investigated, and model selection methods are applied to determine inclusion of prognostic factors. The models are internally validated using bootstrap methods, uniform shrinkage is applied to correct for optimism, and the optimism adjusted models are compared in terms of prognostic factor associations, baseline hazard estimates, and individual prediction estimates.

In **Chapter 5** both cause-specific and subdistribution approaches are applied to develop and internally validate flexible parametric prognostic models to predict the risk of antenatal adverse events in pregnancies with early-onset pre-eclampsia, incorporating delivery of the baby as a competing event. Multiple imputation methods are implemented, fractional polynomial associations are investigated, and model

selection methods are applied. The models are internally validated using bootstrap methods, and uniform shrinkage is applied to correct the subdistribution model for optimism. The models are compared in terms of prognostic factor associations, baseline hazard estimates, and individual prediction estimates.

In **Chapter 6** the flexible parametric prognostic model developed in Chapter 4, and the flexible parametric subdistribution model developed in Chapter 5, are externally validated using an independent dataset of early-onset pre-eclampsia pregnancies. The resulting models, and their measures of prognostic ability, are compared.

Chapter 7 describes the design and results of a simulation study to assess the effects of competing events on measures of prognostic performance. The level of miscalibration introduced when assessing overall calibration and not appropriately accounting for competing events is investigated for different proportions of the event of interest and the competing event. The chapter concludes with recommendations regarding which circumstances competing risks methodology significantly alters measures of prognostic ability, and thus should be applied, in prognostic model research.

Finally, **Chapter 8** concludes the thesis with a summary and discussion of key findings and recommendations regarding when and how competing risks should be accounted for in prognostic model research. The limitations of the research are discussed, and recommendations for future research are provided.

2 AN EMPIRICAL EVALUATION OF THE PRESENCE AND REPORTING OF COMPETING EVENTS IN SYSTEMATIC REVIEWS OF PREDICTION MODEL STUDIES

2.1 Introduction

In order for a prediction model to be useful, the model's estimates of an individual's risk of experiencing a particular outcome in the future need to be as accurate as possible. If competing events are present but not correctly accounted for, the estimated risk predictions provided by the model may not accurately reflect the individual's real-world risk. The inflation in absolute risk estimates resulting from not accounting for competing events is referred to as "*competing risks bias*" (Schatzkin and Slud, 1989, Schumacher et al., 2016, Walraven and McAlister, 2016).

The presence of competing risks bias in prediction model research could result in biased information being used to inform treatment decisions. A study comparing prediction models developed using standard time-to-event methods with a model developed using competing risks methods found an additional 10% of women were classified as high risk for developing coronary heart disease when the competing events were not correctly accounted for (Wolbers et al., 2009). Logically, the amount of bias in absolute risk estimates is associated with the number of observed competing events, more events lead to greater bias. Walraven and McAlister developed an (unpublished) model to assess relative bias in Kaplan-Meier based risk estimates compared to the "true" (competing risks adjusted) values (Walraven and McAlister, 2016). The model was applied to 16 studies and estimated the median relative increase in absolute risk estimates due to competing risks bias, which was 5.7% (range: 1.2% to 65.8%) (Walraven and McAlister, 2016).

Furthermore, systematic reviews of prediction model studies are increasing, and are being used to synthesise the evidence from primary studies and guide the quality of evidence to support existing prediction models (Debray et al., 2017). Such reviews need to consider whether competing events are present and appropriately managed within the primary studies, to enable a thorough assessment of the quality and risk of bias in the primary studies (Wolff et al., 2019). To the author's knowledge, there has been no previous evaluation of the handling of competing events specifically within systematic reviews of prediction models.

2.1.1 Aims

The aim of this chapter is to empirically investigate the presence and reporting of competing events, and competing risks bias, in published systematic reviews of prediction model studies. Therefore, this evaluation of systematic reviews will investigate:

- Whether systematic reviews of prediction model studies contain prediction models with high potential for competing risks bias; and
- Whether these biases are reported in the systematic review articles and/or acknowledged during quality assessment of the included prediction model studies.

2.2 Methods: Evaluation of systematic reviews

2.2.1 Strategy for searching and selecting relevant systematic review articles

To identify systematic review articles of prediction model studies, an extensive search strategy (Ingui and Rogers, 2001), which includes a broad range of relevant search terms and is shown to have a high level of sensitivity (98.2%), was applied. The search strategy was refined by specifying the article title must include either the term "systematic review" or "meta-analysis", as suggested in the PRISMA[i] statement (Moher et al., 2009). This adaptation was unlikely to significantly affect the sensitivity of the search strategy; a recent review found 94% (239/255) of non-Cochrane[ii] systematic review articles include one of these terms in their title (Page et al., 2016).

The MEDLINE database (through Pubmed) was used to search for articles published after the 1st January 2015 to the search date (3rd April 2017) and was restricted to human studies. Based on testing searches, it was anticipated that this date range would result in approximately 30 systematic review articles being identified. After discussion with the research team, 30 articles was considered sufficient to provide an overview of the evaluation of competing risks bias within the current literature on systematic reviews of prediction model studies. The detailed search strategy for the evaluation is listed in Table 2.1.

[i] The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement is an evidence-based minimum set of outcomes for reporting systematic reviews and meat-analyses.
[ii] A Cochrane Review is a systematic review of research in health care and health policy that is published in the Cochrane Database of Systematic Reviews.

Table 2.1: Search strategy for identifying systematic reviews of prediction model studies

Database: MEDLINE (Pubmed) on 03 April 2017.	
(1)	Validat*[Title] OR Predict*[Title] OR Rule*[Title]
(2)	Predict*[Title]
(3)	Outcome*[Title] OR Risk*[Title] OR Model*[Title]
(4)	(2) AND (3)
(5)	History[Title] OR Variable*[Title] OR Criteria[Title] OR Scor*[Title] OR Characteristic*[Title] OR Finding*[Title] OR Factor*[Title]
(6)	Predict*[Title] OR Model*[Title] OR Decision*[Title] OR Identif*[Title] OR Prognos*[Title]
(7)	(5) AND (6)
(8)	Decision*[Title]
(9)	Model*[Title] OR Clinical*[Title] OR Logistic Models[MeSH:noexp, Title]
(10)	(8) AND (9)
(11)	Prognostic[Title]
(12)	History[Title] OR Variable*[Title] OR Criteria[Title] OR Scor*[Title] OR Characteristic*[Title] OR Finding*[Title] OR Factor*[Title] OR Model*[Title]
(13)	(11) AND (12)
(14)	(1) OR (4) OR (7) OR (10) OR (13)
(15)	systematic review[Title] OR meta-analysis[Title]
(16)	(14) AND (15)
Limitations: published between 01 January 2015 and 03 April 2017; Humans	

This evaluation focused on systematic review articles which describe the development of prediction models for health-related outcomes. The following definition was selected to describe a prediction model: *"a multivariable (containing two or more variables) statistical model developed to predict an individual's risk of a future outcome over a period of time from a given disease or health state",* adapted from (Collins et al., 2015). This definition allows the inclusion of models which make predictions in healthy individuals from the general population, as well as in clinical populations and those in specified health states, as competing events and competing risks bias can potentially affect these populations too. Diagnostic models, which are used to identify a patient's current disease or health state, are excluded from this definition as competing events are unlikely to prevent outcomes, as these are already present or not. Thus, systematic review articles were only considered suitable for inclusion in the evaluation if they met the following criteria:

1. described a systematic review and/or meta-analysis study; and
2. assessed and/or synthesised two or more primary studies developing prediction models.

These criteria led to the exclusion of systematic review articles that include prediction model research articles which:

1. only identify individual prognostic factors rather than multivariable prediction models; or
2. only validate existing prediction models and do not develop new models; or
3. only assess diagnostic models and not prediction models.

Systematic reviews which included a mixture of prediction model development articles *alongside* prognostic factor, validation, or diagnostic research articles were thus included in the evaluation.

The selection of systematic review articles for this evaluation was determined through a screening of titles and abstracts, followed by (if necessary) an assessment of the full text. Articles that did not meet the above criteria were removed throughout this process, and reasons for exclusion were recorded. All articles were examined by the first reviewer (LT), and 5% of the articles were additionally independently examined by a second reviewer (KS or DvdW). Disagreements concerning the inclusion of individual articles were addressed via discussions between the reviewers (LT, KS, and DvdW).

2.2.2 Data extraction

Information from the included systematic review articles was extracted by the first reviewer (LT) using a data extraction form (Appendix I). A second reviewer (KS) independently extracted information from 20% of the included systematic review articles. Any discrepancies between the data extracted by the two reviewers were

resolved by discussion between the reviewers (LT and KS). Information was extracted in connection to four key items, as outlined below:

2.2.2.1 Item 1: What were the characteristics of each systematic review?

This item aimed to summarise the characteristics of each included systematic review. The date restrictions (start and end dates) reported in each systematic review's search strategy were recorded. If no start date was reported within the search strategy it was assumed that all research articles published before the end date were included, i.e. the start date was recorded as the inception of the database(s) searched within the review. Information was recorded regarding whether each systematic review aimed to identify prediction model articles which developed models to predict a specific outcome(s) (such as stroke or mortality), or which developed models in populations in a specific health state (such as those in hospital), or indeed both. Finally, both the number of prediction model articles included in each systematic review and the reported number of prediction models was recorded. Some systematic reviews identified prognostic factor, validation, and diagnostic research articles *as well as* development articles. Additionally, some identified articles which developed *multiple* prediction models; thus the number of prediction model articles and reported number of models in each systematic review may not be equal.

2.2.2.2 Item 2: What is the potential for competing risks bias affecting each systematic review?

The reported characteristics of each prognostic model developed within each prognostic model study included in each systematic review were assessed using a combination of criteria. These criteria, utilised in published articles addressing competing risks bias (Koller et al., 2012, Schumacher et al., 2016, Walraven and McAlister, 2016), were considered to assess the risk of competing risks bias within each systematic review (further details are provided in the Discussion). The three

criteria are outlined below, followed by a description of the process used to assess the risk of competing risks bias within each systematic review.

Criterion for competing risks bias 1: The prediction model investigates outcomes other than all-cause mortality

It is generally accepted that outcomes such as all-cause mortality, and composite outcomes which contain all-cause mortality including progression-free survival, are not at risk of competing events (Koller et al., 2012, Schumacher et al., 2016, Walraven and McAlister, 2016), as nothing can prevent these outcomes from occurring. However, death is often considered to be a competing event to a number of other outcomes, including: non-fatal disease-specific outcomes (such as antenatal adverse events in pre-eclampsia); and cause-specific death (such as death from cancer), as death from any other cause prevent these outcomes from occurring. The reported primary outcome(s) of interest of each prediction model identified within each systematic review were recorded. Systematic reviews were considered to meet this criterion if they included prognostic models which predicted outcomes other than all-cause mortality.

Criterion for competing risks bias 2: The baseline population contains frail and/or elderly populations

It is also generally accepted that frail and elderly populations have an increased risk of experiencing competing events (Koller et al., 2012). Disease and health states indicating increased morbidity, chronic diseases, and severe or critical illnesses (such as populations with cardiovascular disease, in intensive care, or receiving cancer therapies) were reasoned to indicate frail populations; as these populations have a greater likelihood of experiencing competing events (such as death) (Koller et al., 2012). Elderly populations are considered susceptible to competing risks bias due to increased disease accumulation and frailty (Koller et al., 2012), thus the populations have a greater likelihood of experiencing competing events. The reported initial

disease or health state and descriptions of the age of the population which the prediction models identified within each systematic review were developed in were recorded. Systematic reviews were considered to meet this criterion if they contained prediction models developed in elderly or frail populations.

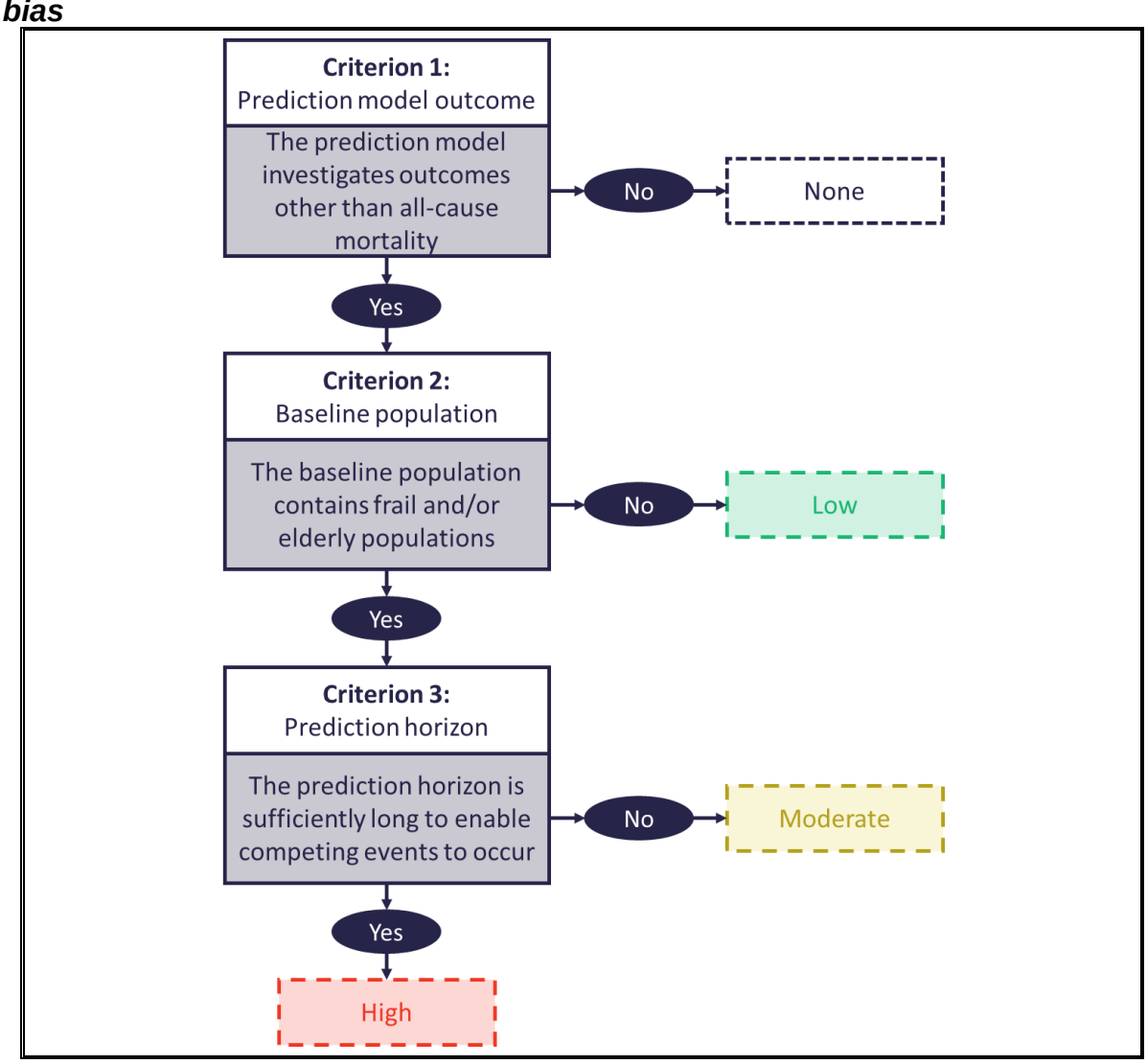
Criterion for competing risk bias 3: The prediction horizon is sufficiently long to enable competing events to occur

A prediction horizon is the time for which an individual's risk is assessed; for example a model which predicts the ten year risk of coronary heart disease (Wolbers et al., 2009) has a ten year prediction horizon. A longer prediction horizon enables a greater number of competing events to occur. A summary of the prediction horizons of the prediction models identified within each systematic review were recorded. Where this information was not reported, a summary of follow-up time was recorded as an alternative. Systematic reviews which included prediction models with prediction horizons of a year or greater were considered to meet this criterion. The cut off at one year was a pragmatic decision for the purpose of the thesis, ideally clinical expertise should be sought to determine a sensible cut-point for this assessment based on the studies health state and outcome of interest.

Overall assessment of the potential for competing risks bias within each systematic review

The above criteria were evaluated in turn within each systematic review; the process is summarised in Figure 2.1.

Figure 2.1: Classification system for assessing the potential for competing risks bias



Systematic reviews which *only* identified prediction models with all-cause mortality (or composites containing all-cause mortality) outcomes were not considered susceptible to competing risks bias and were thus classified as no potential for competing risks bias. Those systematic reviews which identified prediction models with other outcomes (Criterion 1) were evaluated further. Systematic reviews which *only* identified prediction models developed in young and healthy populations were considered to have a small chance of competing risks bias, and were thus classified as low potential for competing risks bias. Those systematic reviews which identified prediction models developed in frail and/or elderly populations (containing participants

over 60 years of age[i], Criterion 2) were evaluated further. Finally, systematic reviews which *only* identified prediction models with prediction horizons less than one year (considered to be sufficiently short to reduce the likelihood of the occurrence of competing events) were classified as moderate potential for competing risks bias; Whereas those systematic reviews which identified prediction models with prediction horizons one year or greater (Criterion 3) were classified as high potential for competing risks bias.

Systematic reviews classified as high potential are likely to contain a large proportion of prediction model studies with competing events. Whereas those classified as no potential are unlikely to contain any prediction model studies with competing events. However, a prediction model study with competing events is not necessarily susceptible to competing risks bias. If the prediction model study appropriately accounts for the competing events, for example through the use of competing risks regression (as outlined in Chapter 1), the results of the study will not be biased. However, it is suspected by the reviewers that the appropriate methods are scarcely applied in prediction model research, thus the presence of competing events is likely to be a strong indicator for the presence of competing risks bias.

2.2.2.3 Item 3: Were competing events addressed in the systematic review article?

This item aimed to summarise whether any of the included systematic review articles directly addressed issues related to competing risks bias. To do this, a search of key terms related to competing events (listed in Table 2.2) was conducted in each systematic review using the *Adobe Acrobat Reader DC Find* function. Where key terms were identified within the systematic review article, the associated text surrounding the term was extracted and reported. Additionally, as competing risks bias is connected to

[i] 60 is acknowledged as an artificial cut-off point for “elderly” used by the World Health Organisation ORGANIZATION, W. H. 1994. *Health care for the elderly: a manual for primary health care workers*.

overestimation of Kaplan-Meier estimates, each systematic review was examined for reference to, and/or depiction of, Kaplan-Meier curves.

Table 2.2: Key terms related to competing events

Key terms
competing risk(s); competing event(s); competing cause(s); competing bias; cause specific; cause-specific; subdistribution; sub-distribution; cumulative incidence; Fine & Gray; Fine and Gray;

2.2.2.4 Item 4: Was competing risks part of the quality assessment performed within the systematic review?

Information was extracted regarding whether each systematic review reported performing a quality assessment of the included prognostic model studies. If done, the name of the quality assessment tool implemented to perform the assessment was recorded. Each quality assessment tool identified was obtained (where possible) and examined to determine whether it explicitly mentioned competing risks and the potential for competing risk bias.

2.2.3 Analysis methods

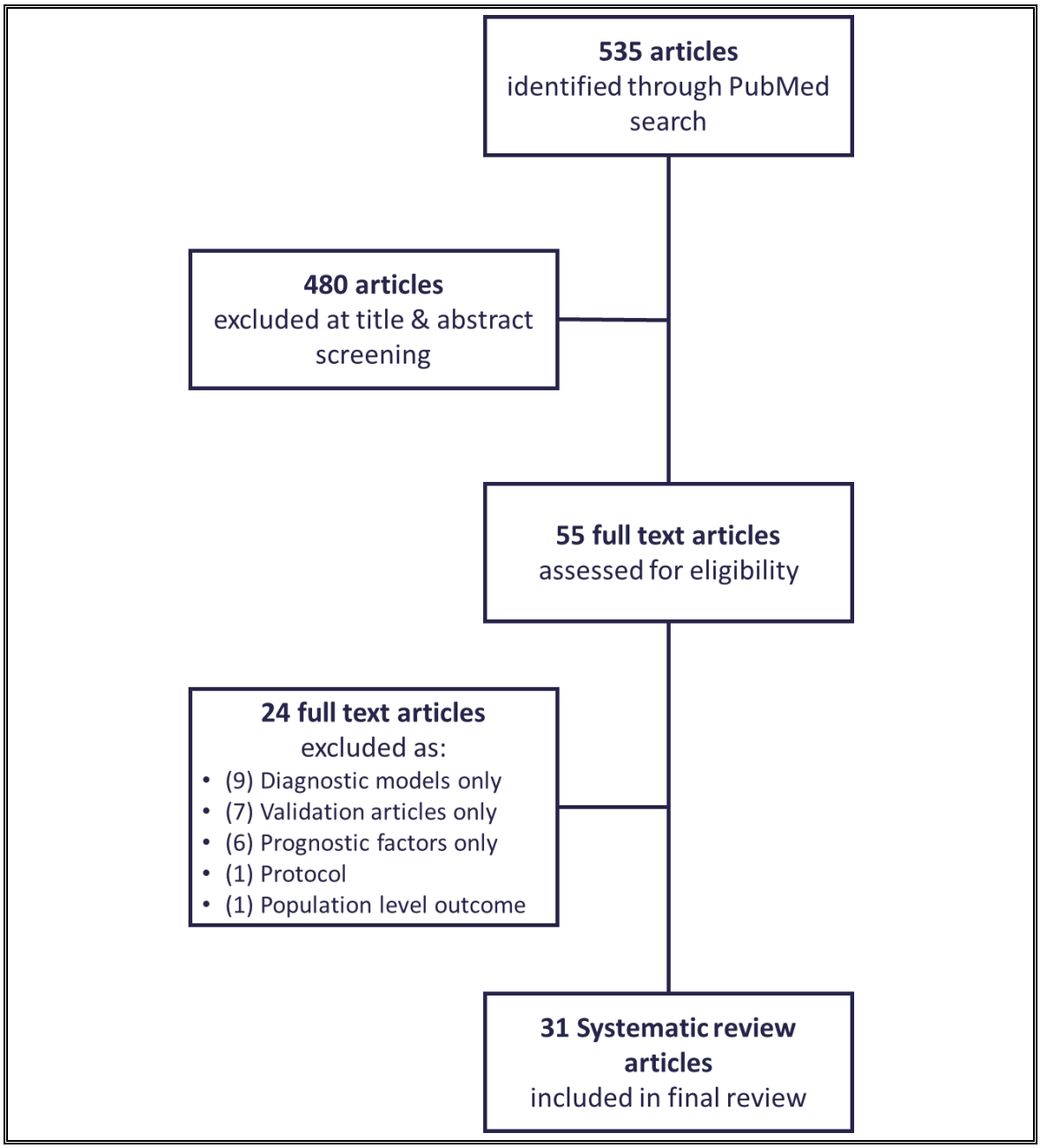
A narrative synthesis of the information extracted from the systematic review articles was conducted, the preliminary synthesis consisted of tabulation and textual descriptions of the extracted information.

2.3 Results: Evaluation of systematic reviews

2.3.1 Search and selection of relevant systematic reviews

A flow diagram depicting the selection process for this evaluation is given in Figure 2.2.

Figure 2.2: Flow diagram of systematic review article selection process for this evaluation



The initial PubMed search identified 535 articles for consideration. Titles and abstracts were screened to identify potentially relevant articles, leading to 480 irrelevant articles being excluded. Full text manuscripts were retrieved for the

remaining 55 articles; of these 24 did not meet the review selection criteria and were consequently excluded. The main reasons for exclusion at this stage were systematic reviews which: only assessed diagnostic models; only assessed model validation studies; and only assessed prognostic factor studies. The remaining 31 systematic reviews met the evaluation inclusion criteria and were thus considered suitable for the evaluation. Additional reviewers (KS and DvdW) each checked ten randomly selected articles, from the initial 535 identified, against the review inclusion criteria. No discrepancies were identified between reviewer decisions; thus the 31 systematic review articles formed the final set for this evaluation. The findings about items 1 to 4 are now summarised for these 31 articles.

2.3.2 Item 1: What were the characteristics of each systematic review?

The characteristics of the 31 included systematic reviews of prediction model studies are summarised in Appendix II. The majority of the systematic reviews (22, 71.0%) aimed to identify prediction models developed in a specified population, and which predict a pre-specified outcome. Two (6.5%) of the systematic reviews specified *only* the prediction model outcomes, specifically lung cancer (Gray et al., 2016) and predicting future dementia diagnosis (Tang et al., 2015). Seven (22.6%) systematic reviews specified *only* the population in which the prediction models were developed, and not the outcomes predicted by the models. The remaining three articles provided very broad definitions for outcomes predicted by the models, including “behavioural/psychiatric problems” (Linsell et al., 2016), “adverse outcomes” (O'Caoimh et al., 2015), and “late effects” (Salz et al., 2015). All reviews included an evaluation of multiple prediction models, with a median of 16 prediction models per review (range 3 to 363) identified from a median of 20 published articles per review (range 3 to 125). One systematic review (Tang et al., 2015) failed to report the number of models included in the review. It was common for the systematic reviews to identify

more prediction models than model development articles (13, 41.9%), due to primary studies presenting multiple prediction models.

2.3.3 Item 2: What is the potential for competing risks bias affecting each systematic review?

The risk of competing risks bias within each systematic review was assessed by scrutinizing the reported characteristics of the prediction model development studies included within each review. Information pertaining to three criteria (prediction model outcomes, baseline populations, and prediction horizons) were extracted from each systematic review article, these are reported in Table 2.3. Each of the criteria were assessed separately prior to being combined to determine an overall potential for competing risks bias. The results are provided below.

Criterion for competing risk bias 1: The prediction model investigates outcomes other than all-cause mortality

Data which summarise the outcomes of the prediction models identified within each systematic review is displayed in Table 2.3. The 31 systematic reviews identified prediction models with a variety of outcomes, including morbidity (Lim et al., 2015, Marufu et al., 2015, Salz et al., 2015), acute kidney injury (Caragata et al., 2016, Wilson et al., 2016), asthma (Luo et al., 2015, Smit et al., 2015), and venous thromboembolism (Ensor et al., 2016, Mahajerin et al., 2015). Of the 31 systematic reviews, eight (25.8%) reported the inclusion of prediction models with all-cause mortality outcomes, yet only three (9.7%) (Kohn et al., 2015, Oliver et al., 2015, Warnell et al., 2015) included *only* all-cause mortality outcomes. Hence 28 (90.3%) systematic reviews contained prediction models which investigated outcomes other than all-cause mortality and met this criterion.

Table 2.3: The risk of competing risks bias within each systematic review: based on characteristics of the prediction model development studies contained within each systematic review.

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations			Criterion for competing risk bias 3: Prediction horizons	
	Outcomes of the prediction models identified in the review[+]	Do all outcomes contain all-cause mortality?	Disease and health states of prediction model populations	Population age at baseline		Prediction horizon of models identified in the review	
(Ayerbe et al., 2016)	**Coronary artery disease**, myocardial infarction, angina, coronary insufficiency, unstable angina, cardiac death, arrhythmia, revascularisation	No	Chest pain (recent onset)	Not Reported		Range	Discharge to 4 years
(Brunelli and Prefumo, 2015)	**Pre-eclampsia**	No	First trimester pregnancy	Not Reported		Not Reported	
(Caragata et al., 2016)	**Acute kidney injury**	No	Liver transplantation	Not Reported		Range	<48 hours to <10 days
(Damen et al., 2016)	**Cardiovascular Disease**, with coronary heart disease, with stroke, myocardial infarction, and atrial fibrillation, all-cause mortality	No	General population	Range	30 to 74 years	Range	2 to 45 years
(Echouffo-Tcheugui et al., 2015)	**Incident heart failure**	No	General population	Range	18 to 99 years	Range*	2.9 to 14 years
(Ensor et al., 2016)	**Recurrent venous thromboembolism**	No	Cessation of treatment venous thromboembolism	Mean range Median range	52.3 to 53.6 years 54 to 63 years	Range	3 months to 5 years

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations			Criterion for competing risk bias 3: Prediction horizons	
	Outcomes of the prediction models identified in the review[+]	Do all outcomes contain all-cause mortality?	Disease and health states of prediction model populations	Population age at baseline		Prediction horizon of models identified in the review	
(Gray et al., 2016)	**Lung cancer**, death from lung cancer, survival, lung cancer incidence	No	General population	Range	20 to 80 years	Range	1 to 10 years
(Haskins et al., 2015)	Functional outcomes, work related outcomes, pain intensity, recovery, symptom persistence, need for surgical intervention	No	Low Back Pain	Mean range	26 to 56 years	Range	48 hours to 2 years
(Hilkens et al., 2016)	**Intracranial haemorrhage or major bleeding**, major bleeding, intracerebral bleeding, gastrointestinal bleeding	No	Patients on antiplatelet therapy	Mean range	63 to 69 years	Range	2 to 5 years
(Kohn et al., 2015)	**Early all-cause mortality**	Yes	Acute pulmonary embolism	Mean range	56 to 71 years	Max	90 days
(Lim et al., 2015)	Overall morbidity, all-cause, liver failure mortality, overall morbidity and mortality, bile leakage, infection, mortality and hospital stay, liver failure, infection and organ dysfunction, overall morbidity, ascites	No	Patients undergoing liver resection	Not Reported		Max	90 days
(Linsell et al., 2016)	Neurodevelopment outcomes, general behavioural problems, any psychiatric disorders, autism spectrum disorders, attention deficit/hyperactivity disorder	No	Very preterm or very low birth weight infants	Max	32 weeks	Min	18 months
(Luo et al., 2015)	**Asthma development**, multi-trigger wheezing, persistent wheezing, receiving treatment, diagnosis	No	Children	Range	0 to 14 years	Range	1 to 19 years

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations			Criterion for competing risk bias 3: Prediction horizons	
	Outcomes of the prediction models identified in the review[+]	Do all outcomes contain all-cause mortality?	Disease and health states of prediction model populations	Population age at baseline		Prediction horizon of models identified in the review	
(Mahajerin et al., 2015)	**Hospital-associated venous thromboembolism**	No	Pediatric hospital patients	Max	21 years	Not Reported	
(Mao et al., 2015)	**Stroke**	No	Coronary artery bypass grafting	Mean range Median range	62 to 69 years 64.7 to 65.3 years	Range	1 to 30 days
(Marques et al., 2015)	**Osteoporotic fracture**, hip fracture, major osteoporotic fracture, osteoporotic or fragility fracture, clinical vertebral fracture, mortality	No	General population	Range	40 to 100 years	Range	1 to 10 years
(Marufu et al., 2015)	**Mortality, morbidity, mobility**	No	Hip fracture operation	Not Reported		Range	In hospital to more than 1 year
(Meyer et al., 2015)	**Functional outcomes**, Barthel Index or Functional Independence Measure.	No	Patients receiving post-stroke inpatient rehabilitation	Not Reported		Max	Hospital discharge
(O'Caoimh et al., 2015)	Hospitalisation, functional-decline, institutionalisation, death	No	Community-dwelling older adults	Mean range	64.2 to 84.6 years	Range	4.5 months to 9 years
(Oliver et al., 2015)	Mortality	Yes	Emergency laparotomy	Not Reported		Max	30 days
(Quinlivan et al., 2016)	**Repeat self-harm and attempted suicide**, self-harm, self-poisoning	No	Presenting with self-harm or attempted suicide	Not Reported		Range	3 months to 3 years

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations			Criterion for competing risk bias 3: Prediction horizons	
	Outcomes of the prediction models identified in the review[+]	Do all outcomes contain all-cause mortality?	Disease and health states of prediction model populations	Population age at baseline		Prediction horizon of models identified in the review	
(Salz et al., 2015)	Erectile dysfunction, urinary incontinence, arm lymphoedema, psychological morbidity, cardiomyopathy or heart failure, cardiac event, swallowing dysfunction, breast cancer, thyroid cancer.	No	Completing treatment for cancer	Not Reported		Range	6 months to 30 years
(Silver et al., 2015)	**Contrast induced nephropathy**	No	Undergoing procedure using iodinated radiocontrast	Not Reported		Range	2 to 7 days
(Silverberg et al., 2015)	Post-concussion symptom reporting, functional outcome, quality of life, neuropsychological outcomes	No	Mild traumatic brain injury	Range	5 to 80 years	Range	3 to 18 months
(Smit et al., 2015)	**Asthma**	No	Children with asthma-like symptoms	Range	0 to 5 years	Range	1 to 11 years
(Tang et al., 2015)	**Dementia**, Alzheimer's disease	No	Varied, inlcuding general population, elderly, and with diabetes	Range	40 to 99 years	Range	1.5 to 17 years
(Usher-Smith et al., 2016)	**Primary colorectal cancer**, advanced colorectal neoplasia including dysplasia, advanced colorectal neoplasia, colon cancer (distal and proximal), rectal cancer	No	General population	Mean range	36.4 to 67.5 years	Range[*]	10 to 20 years
(Walsh et al., 2016)	**Falls**, multiple or injurious falls	No	Stroke	Mean range	62.0 to 76.4 years	Range[*]	Discharge to 12 months
(Warnell et al., 2015)	**Perioperative mortality**	Yes	Oesophagectomy for cancer in adults	Not Reported		Range	0 to 90 days

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations			Criterion for competing risk bias 3: Prediction horizons	
	Outcomes of the prediction models identified in the review[+]	Do all outcomes contain all-cause mortality?	Disease and health states of prediction model populations	Population age at baseline		Prediction horizon of models identified in the review	
(Williams et al., 2016)	**Colorectal cancer**, with advanced adenoma, with pre-malignant adenomas, with other cancers	No	Symptomatic of colorectal cancer	Range	18 to 89 years	Max	2 years
(Wilson et al., 2016)	**Acute kidney injury**	No	Major non-cardiac surgery	Not Reported		Range	1 week to 30 days

+Text in bold represents the outcome of interest specified by the review.
*** Follow-up measures reported instead of prediction horizon as prediction horizon not reported in the review.**

Criterion for competing risk bias 2: The baseline population contains frail and/or elderly populations

Data which summarise the baseline populations of the prediction models identified within each systematic review is displayed in Table 2.3. The 31 systematic reviews included prediction models developed in a variety of populations. Seven (22.6%) systematic reviews include prediction models developed in populations which have undergone a surgical intervention (Caragata et al., 2016, Lim et al., 2015, Mao et al., 2015, Wilson et al., 2016, Marufu et al., 2015, Oliver et al., 2015, Warnell et al., 2015); six (19.4%) investigate prediction models developed in the general population (Damen et al., 2016, Echouffo-Tcheugui et al., 2015, Gray et al., 2016, Marques et al., 2015, Tang et al., 2015, Usher-Smith et al., 2016); and four (12.9%) include prediction models developed in children (Linsell et al., 2016, Luo et al., 2015, Mahajerin et al., 2015, Smit et al., 2015). Twelve systematic reviews (38.7%) did not contain information regarding the baseline age of the study populations in which the prediction models were developed. However, of the 19 systematic reviews which did report baseline age information, the majority (15, 78.9%) included persons over 60 years of age. Those not including persons over 60 were the four articles investigating outcomes in children. All but two (6.5%) of the systematic reviews (Luo et al., 2015, Smit et al., 2015) included prediction models with baseline populations containing frail and/or elderly populations and thus met this criterion.

Criterion for competing risk bias 3: The prediction horizon is sufficiently long to enable competing events to occur

Data which summarise the prediction horizons of the models identified within each systematic review is displayed in Table 2.3. Where information pertaining to the prediction horizons of the models was not reported, a summary of follow-up time was recorded as an alternative. The systematic reviews included prediction models with varying prediction horizons. Five (16.1%) failed to report prediction horizon information;

Information on the prediction model follow-up time was used as a proxy measure in three (9.7%) instances (Echouffo-Tcheugui et al., 2015, Usher-Smith et al., 2016, Walsh et al., 2016). Prediction models with prediction horizons of one year or greater were observed in 20 (64.5%) of the systematic reviews, hence met this criterion. Further, nine (29.0%) systematic reviews included prediction models with horizons which were ten years or longer. The greatest reported prediction horizon was 45 years (Damen et al., 2016).

Overall assessment of the potential for competing risks bias within each systematic review

Based on the three criteria for competing risks bias, an overview of the final assessment of the potential for competing risks bias is provided in Table 2.4. Of the 31 systematic reviews, three (9.7%) *only* included prediction models with all-cause mortality outcomes (Criterion 1 not present). It is unlikely that competing events would be present in the prediction model development studies, and thus would be unlikely to cause bias in the conclusions of the systematic review. Therefore, these three systematic reviews were classified as having no potential for competing risks bias. Two (6.5%) of the systematic reviews included prediction models with outcomes other than all-cause mortality (Criterion 1 present), but which were developed in young and healthy populations (Criterion 2 not present). As competing events (such as death) occur less frequently in young, healthy populations, the potential for bias in the conclusions of the systematic review are small. These two systematic reviews were classified as having low potential for competing risks bias. Seven (22.6%) of the systematic reviews included prediction models with outcomes other than all-cause mortality (Criterion 1 present), which were developed in elderly or frail populations (Criterion 2 present), but with prediction time horizons less than 1 year (Criterion 3 not present). As shorter prediction horizons restrict large numbers of competing events from occurring, the potential for bias in the conclusions of the systematic review was

considered moderate. These seven systematic reviews were classified as having intermediate potential for competing risks bias. Finally, the majority (19, 61.3%) of the systematic review articles met all three criteria, and were thus classified as high potential for competing risks bias. All of the systematic reviews which included prediction models developed in the general population (Damen et al., 2016, Echouffo-Tcheugui et al., 2015, Gray et al., 2016, Marques et al., 2015, Tang et al., 2015, Usher-Smith et al., 2016) were categorised as high potential for competing risks bias.

The 19 systematic reviews classified as high potential for competing risks bias are expected to contain prediction model studies in which competing events were likely to be present. Those nine systematic reviews classified as either low or intermediate may also contain prediction model studies with competing events, though the chances are reduced. This suggests that competing events are a major issue to consider in systematic reviews of prediction model research, as 26 out of 31 systematic reviews were classified moderate to high potential for competing risk bias being an issue. The presence of competing events is likely to cause bias in the conclusions made by the systematic reviews, if not acknowledged or appropriately accounted for. Hence, guidance is needed to prompt those undertaking systematic reviews to examine competing risks and biases within the prediction model studies included in the review.

Table 2.4: An assessment of the potential for competing risk bias in each systematic review

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
	Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Kohn et al., 2015)	No	*Only* all-cause mortality outcomes.					None
(Oliver et al., 2015)	No	*Only* all-cause mortality outcomes.					None
(Warnell et al., 2015)	No	*Only* all-cause mortality outcomes.					None
(Luo et al., 2015)	Yes	Includes other outcomes.	No	Children in general population have low risk of frailty.			Low
(Smit et al., 2015)	Yes	Includes other outcomes.	No	Children in general population have low risk of frailty			Low
(Caragata et al., 2016)	Yes	Includes other outcomes.	Yes	Transplantation indicative of critical illness.	No	Short prediction horizon makes competing events unlikely.	Moderate
(Lim et al., 2015)	Yes	Includes other outcomes.	Yes	Liver rescetion indicative of frail population.	No	Short prediction horizon makes competing events unlikely.	Moderate

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
	Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Mao et al., 2015)	Yes	Includes other outcomes.	Yes	Coronary artery bypass grafting indicative of frail population.	No	Short prediction horizon makes competing events unlikely.	Moderate
(Meyer et al., 2015)	Yes	Includes other outcomes.	Yes	Stroke indicative of frail population.	No	Short prediction horizon makes competing events unlikely.	Moderate
(Silver et al., 2015)	Yes	Includes other outcomes.	Yes	Investigations using iodinated radiocontrast indicative of frail populations.	No	Short prediction horizon makes competing events unlikely.	Moderate
(Brunelli and Prefumo, 2015)	Yes	Includes other outcomes.	Yes	First trimester pregnancies indicative of frail populations.	Not reported	Insufficient information to assess prediction horizon.	Moderate
(Mahajerin et al., 2015)	Yes	Includes other outcomes.	Yes	Patients in pediactric hospital indicative of frail population.	Not reported	Insufficient information to assess prediction horizon.	Moderate
(Ayerbe et al., 2016)	Yes	Includes other outcomes.	Yes	Chect pain indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
	Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Damen et al., 2016)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Echouffo-Tcheugui et al., 2015)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Ensor et al., 2016)	Yes	Includes other outcomes.	Yes	Venous thromboembolism indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Gray et al., 2016)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Haskins et al., 2015)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Hilkens et al., 2016)	Yes	Includes other outcomes.	Yes	Antiplatelet therapy indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
	Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Linsell et al., 2016)	Yes	Includes other outcomes.	Yes	Very preterm & low birth weight infants indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Marques et al., 2015)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Marufu et al., 2015)	Yes	Includes other outcomes.	Yes	Hip fractures indicative of elderly populations with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(O'Caoimh et al., 2015)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Quinlivan et al., 2016)	Yes	Includes other outcomes.	Yes	Self-harm & suicide indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Salz et al., 2015)	Yes	Includes other outcomes.	Yes	Cancer indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High

Systematic review reference	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
	Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Silverberg et al., 2015)	Yes	Includes other outcomes.	Yes	Traumatic brain injury indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Tang et al., 2015)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Usher-Smith et al., 2016)	Yes	Includes other outcomes.	Yes	Population includes elderly participants with increased frailty.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Walsh et al., 2016)	Yes	Includes other outcomes.	Yes	Stroke indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Williams et al., 2016)	Yes	Includes other outcomes.	Yes	Cancer indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High
(Wilson et al., 2016)	Yes	Includes other outcomes.	Yes	Major surgery indicative of frail population.	Yes	Prediction horizon sufficiently long for competing events to occur.	High

2.3.4 Item 3: Were competing events reported in the systematic review article?

Despite the large potential for competing risks bias in most of the systematic review articles, only a few directly addressed issues related to competing events and their associated biases. Two (6.5%) of the included systematic review articles specifically mentioned the terms “competing risk” or “competing event” and both were considered to have high potential for competing risks bias. The first considered the presence of competing events (specifically deaths), and concluded the proportion of events *“was likely to be very small… and therefore the model predictions would not change importantly if a competing risks model had been used”* (Ensor et al., 2016). The other included a prediction model developed using competing risks regression methods (Damen et al., 2016). However, the term “competing risks” only appeared in an online supplementary table, and not in the published article. A single (3.2%) systematic review article included other key terms associated with competing events; the term “cumulative incidence” was used throughout the article to refer to the proportions of the prediction model study populations that had asthma (Smit et al., 2015), and was thus not used in a competing risks context. Finally, three (9.7%) systematic reviews included or discussed Kaplan-Meier curves to display risk groups as a method for assessing model calibration (Ensor et al., 2016, Salz et al., 2015, Walsh et al., 2016). All three were considered to have high potential for competing risks bias, and so Kaplan-Meier curves are potentially misleading.

2.3.5 Item 4: Was competing risks part of the quality assessment performed within the systematic review?

The majority (29, 93.5%) of the systematic reviews reported performing a quality assessment of the included prediction model studies. A range of quality assessment tools were reported, details are given in Table 2.5.

Table 2.5: Quality assessment tools from systematic review articles

Systematic review reference	Quality Assessment Tool
Ayerbe et al., 2016	QUADAS (Whiting et al., 2003)
Brunelli and Prefumo, 2015	Collins criteria (Collins et al., 2011, Collins et al., 2013)
Caragata et al., 2016	None reported
Damen et al., 2016	CHARMS (Moons et al., 2014)
Echouffo-Tcheugui et al., 2015	Authors own
Ensor et al., 2016	PROBAST (Wolff et al., 2019)
Gray et al., 2016	Authors own
Haskins et al., 2015	QUIPS (Hayden et al., 2013) & Authors own
Hilkens et al., 2016	CHARMS (Moons et al., 2014)
Kohn et al., 2015	QUADAS-2 (Whiting et al., 2011)
Lim et al., 2015	QUIPS (Hayden et al., 2013)
Linsell et al., 2016	QUIPS (Hayden et al., 2013)
Luo et al., 2015	CASP (2017)
Mahajerin et al., 2015	None reported
Mao et al., 2015	Authors own
Marques et al., 2015	QUADAS (Whiting et al., 2003)
Marufu et al., 2015	Altman framework (Altman, 2001)
Meyer et al., 2015	QUIPS (Hayden et al., 2013)
O'Caoimh et al., 2015	QUIPS (Hayden et al., 2013)
Oliver et al., 2015	Altman framework (Altman, 2001)
Quinlivan et al., 2016	QUADAS (Whiting et al., 2003) & STARD (Bossuyt et al., 2003)
Salz et al., 2015	Authors own
Silver et al., 2015	QUIPS (Hayden et al., 2013)
Silverberg et al., 2015	Authors own
Smit et al., 2015	CHARMS (Moons et al., 2014)
Tang et al., 2015	Newcastle-Ottowa scale (GA Wells, 2014)
Usher-Smith et al., 2016	TRIPOD (Collins et al., 2015)
Walsh et al., 2016	McGinn criteria (McGinn et al., 2000)
Warnell et al., 2015	QUIPS (Hayden et al., 2013)
Williams et al., 2016	CASP (2017)
Wilson et al., 2016	TRIPOD (Collins et al., 2015)

The most commonly used tool for assessment of the prediction model development studies was the Quality in Prognosis Studies (QUIPS) tool (Hayden et al., 2013), which was reportedly utilised in seven (22.6%) of the systematic reviews. This was followed by the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool (Whiting et al., 2003, Whiting et al., 2011), used in four (12.9%) of the systematic reviews. A number of authors (6, 19.4%) developed their own tools or criteria for quality assessment, whereas four others (12.9%) replicated or amended tools employed in similar reviews (Altman, 2001, Collins et al., 2011, Collins et al.,

2013, McGinn et al., 2000). Two (6.5%) systematic review articles did not report performing a quality assessment of the included prediction model studies.

Each of the listed quality assessment tools was examined to determine whether the assessment of competing events was directly advised. The PROBAST tool (Wolff et al., 2019), explicitly mentions competing risks in the context of the analysis of complexities in the data. None of the other tools explicitly refer to competing events and the associated biases. This may be because the other tools are not specific to prediction model research, and rather focus on prognostic factor research (QUIPS) and diagnostic tests (QUADAS). The PROBAST tool was rarely used in the systematic review articles identified as it was only formally released in 2019.

2.4 Discussion

In this chapter, an evaluation of the presence and reporting of competing events in 31 systematic reviews of prediction model studies was conducted. A classification system for the risk of competing risks bias was developed and applied to each systematic review. The key findings and conclusions of this evaluation are summarised in Box 2.1 and discussed below.

Box 2.1: Key findings

1. Competing events are often present in the prediction models contained within systematic reviews of prediction model studies.
2. Systematic reviews of prediction model studies rarely report the presence or assessment of competing risks.
3. A wide variety of quality assessment tools are utilised in systematic reviews of prediction models studies, few consider competing risks.

2.4.1 Key findings

The key findings of this evaluation are presented and discussed in more detail below:

2.4.1.1 Competing events are often present in the prediction models contained within systematic reviews of prediction model studies.

Of the 31 systematic reviews evaluated, 90.3% were found to include prediction models with outcomes other than all-cause mortality, making them susceptible to competing events. Further, 61.3% were classified as having high potential for competing risks bias, when the prediction model outcome, baseline population, and prediction horizon were considered. This suggests that competing events are commonly present in prediction model development studies, and competing risks bias will often be a potential concern for systematic reviews of these studies. Consequently, the competing events should be acknowledged and examined in systematic reviews of prediction model studies.

2.4.1.2 Systematic reviews of prediction model studies rarely report the presence or assessment of competing risks.

Despite the high potential for competing events in prediction model studies, few systematic reviews reported or assessed issues related to competing risks in the published articles. Only two (6.5%) of the systematic review articles evaluated directly reported on competing risks. If the statistical methods used to develop prediction models are not appropriate for the complexities of the data, the estimated predictive performance of the model may be biased (Moons et al., 2015). Not reporting the presence of competing risks could result in biases in the conclusions made by the systematic reviews.

2.4.1.3 A wide variety of quality assessment tools are utilised in systematic reviews of prediction models studies, but few consider competing risks

This evaluation identified a number of quality assessment tools currently being used to assess prediction model studies in systematic reviews. The array and inconsistency of quality assessment tools used by the included systematic reviews highlights the absence of agreed quality criteria for assessing the risk of bias in prognostic modelling studies. However, the Prediction study Risk Of Bias Assessment Tool (PROBAST) has recently been developed to address this issue. This was the only quality assessment tool used by the systematic reviews which explicitly referred to competing events. The tool directly refers to competing risks when considering the risk of bias in the analysis of the prediction mode study: *"Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?"* (Wolff et al., 2019). This tool was released in 2019, and so it hopefully will help to improve the assessment of competing risks bias in future systematic reviews of prediction model studies.

2.4.2 Limitations and further research

To the author's knowledge, this is the largest evaluation of competing events in systematic reviews of prediction model studies. The evaluation highlights the high presence of competing events in prediction model research, motivating the need to consider competing events in systematic reviews which aim to evaluate such studies. Nevertheless, the evaluation does not encompass all systematic reviews of prediction model studies. The systematic reviews which were not considered, due to not meeting inclusion criteria, may differ to those evaluated. For example, inclusion of studies containing the terms "systematic review" or "meta-analysis" in the article title is likely to identify systematic review conforming to the PRISMA reporting guidelines. However, reviews which follow reporting guidelines may be of better quality, or at least report on more key items, than those which do not.

The classification system used to assess the potential for competing risk bias was developed by combining criteria from articles which similarly aimed to assess the presence of competing risks bias in other settings (Austin and Fine, 2017, Koller et al., 2012, Schumacher et al., 2016, Walraven and McAlister, 2016). Criteria thought to be applicable to prediction model studies, and likely to be reported in the systematic review articles, were selected. Criteria discussed in the articles but not included in the classification system include: the reporting of Kaplan-Meier estimates and assessing the number and proportion of competing events. These were not included as criteria in the classification system as it was thought these would not typically be reported in systematic review articles. The tool was developed to pragmatically determine whether competing events were likely to be present in the included prediction model studies, as it was considered beyond the scope of the thesis to obtain individual study data for all included studies to determine whether competing events were truly present. Further, the presence of competing events in a prediction model study does not necessarily imply that the study contains competing risks bias. If a study identifies the competing events and appropriately accounts for these, for example, through competing risks

regression analysis, then no such bias will be observed. It is suspected by the author that appropriate methods are rarely utilised in prediction model research, and thus the presence of competing events is likely to be indicative of competing risks bias. Thus further research to determine whether competing events were indeed present, and investigating the management of these competing events, is required.

The following chapter describes a review which investigates the presence, reporting, and management of competing events in prediction model studies classified as high potential for competing risks bias, i.e. the most susceptible to competing risks bias.

3 A REVIEW INTO THE PRESENCE, REPORTING AND MANAGEMENT OF COMPETING EVENTS IN PREDICTION MODEL DEVELOPMENT STUDIES

3.1 Introduction

The previous chapter established that competing events and their associated biases are often a potential concern in systematic reviews of prediction model studies; yet these are rarely considered. To emphasise why the issue of competing risks bias should be addressed in future systematic reviews, it is necessary to further investigate the magnitude of competing events and their biases at the primary study (rather than systematic review) level, i.e. within prediction model development studies. Therefore, this chapter provides a detailed assessment of the presence, reporting, and management of competing events in prediction model development studies.

3.1.1 Aims

The aim of this chapter is to review how prediction model development studies handle competing events in clinical settings where competing events are a likely issue. Specifically, this review will:

- Investigate the potential for competing risks bias in each individual prediction model developed within the prediction model studies.
- Explore how competing events are managed within prediction model development studies.
- Determine whether competing events, and their associated biases, are reported in the published prediction model study articles.

3.2 Methods: Review of prediction model development studies

3.2.1 Strategy for searching and selecting relevant prediction model study articles

This review focused on published prediction model study articles which describe the development of prediction models *in clinical settings where competing events are a likely issue*. The prediction model development studies were identified from a subset of systematic review articles identified in Chapter 2. Initially, the systematic reviews examined in Chapter 2 were screened to ascertain which were likely to contain prediction models affected by competing events. Consequently, only those 19 systematic reviews categorised with a high potential for competing risk bias were considered. An additional list of diseases and population characteristics, suggested to identify populations which are susceptible to competing events (Koller et al., 2012), presented in Table 3.1, was applied to further reduce the subset of systematic reviews.

Table 3.1: List of diseases and populations susceptible to competing events

Diseases and population characteristics
Atrial fibrillation; cardiac failure; coronary heart disease; stroke; aneurysm; prostate cancer; colorectal cancer; breast cancer; chronic leukaemia; cancer screening; critical care; transplant; chronic obstructive pulmonary disease; elderly patients (aged 65+ years).

The full-text articles of all prediction model studies examined within the selected systematic reviews were obtained. Prediction model studies were only considered suitable for inclusion to this review if they met the following criteria:

1. Described the development of a prediction model; and
2. Applied time-to-event analysis methods to develop the prediction model; and
3. Had a clinical setting which reflects *at least one* of the disease and population characteristics listed in Table 3.1.

These criteria led to the exclusion of prediction model studies which:

1. *Only* identified individual prognostic factors without combining to develop a prediction model; or
2. *Only* validated existing prediction models without developing new models; or
3. *Only* developed diagnostic, and not prediction, models; or
4. *Only* used analysis methods other than time-to-event regression, to develop prediction models; or
5. Had a clinical setting not considered to be susceptible to competing events.

Conference abstracts were excluded as they contained insufficient detail of the prediction model study for meaningful review.

3.2.2 Data extraction

Relevant information from the included prediction model studies was extracted from the published articles and compiled by the first reviewer (LT) using a data extraction form (Appendix III). A second reviewer (KS) independently extracted information from a third of the included prediction model studies. Any discrepancies between the data extracted by the two reviewers were resolved by discussion between the reviewers (LT and KS). Information was extracted and then narratively summarised in regards to four key items, as outlined below:

3.2.2.1 Item 1: What were the characteristics of each prediction model study?

Information regarding the number of individual prediction models developed within each prediction model study was recorded, as well as the source of the study data (e.g. randomised trial, cohort, or nested case-control) and the total number of participants included in the study.

3.2.2.2 Item 2: What is the potential for competing risks bias affecting each individual prediction model?

The reported characteristics of each individual prediction model developed within each prognostic model study were examined using the same criteria and classification system as developed in Chapter 2. This classification system evaluates the prediction model outcome, baseline population, and prediction horizon in turn and combines the results using the process summarised in Figure 2.1. Again, where specific prediction horizons were not reported, the maximum reported follow-up time was recorded as an alternative.

A list of potential competing events likely to prevent the prediction model outcome from occurring was compiled for each prediction model study. Mortality was considered a potential competing event for *all* individual prediction models predicting outcomes other than all-cause mortality. Potential competing events were additionally determined through examination of the published prediction model study articles for the mention of any events likely to prevent the prediction model events. Further, potential competing events listed for individual prediction models with similar outcomes were compared and added to.

3.2.2.3 Item 3: Were competing events reported in the published prediction model study articles?

The reporting of competing events was examined for the individual prediction models considered to have potential for competing risks bias (classified as low, moderate, or high in Item 2). Information was extracted on the following:

1. The number of prediction model events;
2. The number of reported competing events;
3. The proportion of all reported events which were competing events;
4. The presence of key terms related to competing events (as listed in Table 2.2);
5. Whether Kaplan-Meier curves are presented or discussed;
6. The number of prognostic factors included in the final prediction models; and
7. Any prognostic factors considered to be associated with the competing events.

It has been demonstrated that the level of competing risks bias is strongly associated with the proportion of all observed outcomes that are competing events (Berry et al., 2010, Schumacher et al., 2016, van Walraven and Hawken, 2016, Wolkewitz et al., 2014). Thus this information was extracted. A search of key terms related to competing events (Table 2.2) was conducted using the Adobe Acrobat Reader DC Find function. References to, and depictions of, Kaplan-Meier curves were investigated, as these estimates of absolute risk over time are known to be inflated when competing events are present. Finally, the number of prognostic factors included in the final models, as well as whether these are likely to be associated with competing events, were considered. It has been demonstrated that associations between predictors (such as prognostic factors) and outcomes can change importantly when competing events are appropriately accounted for (Berry et al., 2010, Dignam et al., 2012, Schatzkin and Slud, 1989, Wolkewitz et al., 2014). Appropriately accounting for competing events can alter the magnitude, and in some cases the direction, of the estimated association, particularly when the predictors are strongly associated with the competing event (Berry et al., 2010, Schatzkin and Slud, 1989). Prognostic factors shown to be associated with mortality (anticipated to be a common potential competing event) include age and numerous comorbidities; defined in this instance as chronic diseases or disorders listed in the Charlson comorbidity index (Charlson et al., 1987), listed in Table 3.2.

Table 3.2: Comorbidities listed in the Charlson comorbidity index

Conditions included in Charlson comorbidity index
Myocaridal infarct, Congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disorder, ulcer disease, mild liver disease, diabetes, hemiplegia, moderate or severe renal disease, diabetes with end organ damage, any tumour, leukemia, lymphoma, modersate or severe liver disease, matastatic solid tumor, AIDS.

3.2.2.4 Item 4: How were competing events managed in each prediction model study?

Information regarding how competing events were managed within each prediction model study was examined. Specifically:

1. Whether statistical regression methods which appropriately account for competing events, were employed to develop the individual prediction models;
2. Whether participants who experienced competing events were excluded from the study;
3. Whether participants who experienced competing events were censored at the point of experiencing the competing events;
4. whether participants who experienced competing events were managed in any other way;
5. Whether the study also included the validation of the individual prediction models; and
6. How competing events were managed during the validation process.

3.2.3 Analysis methods

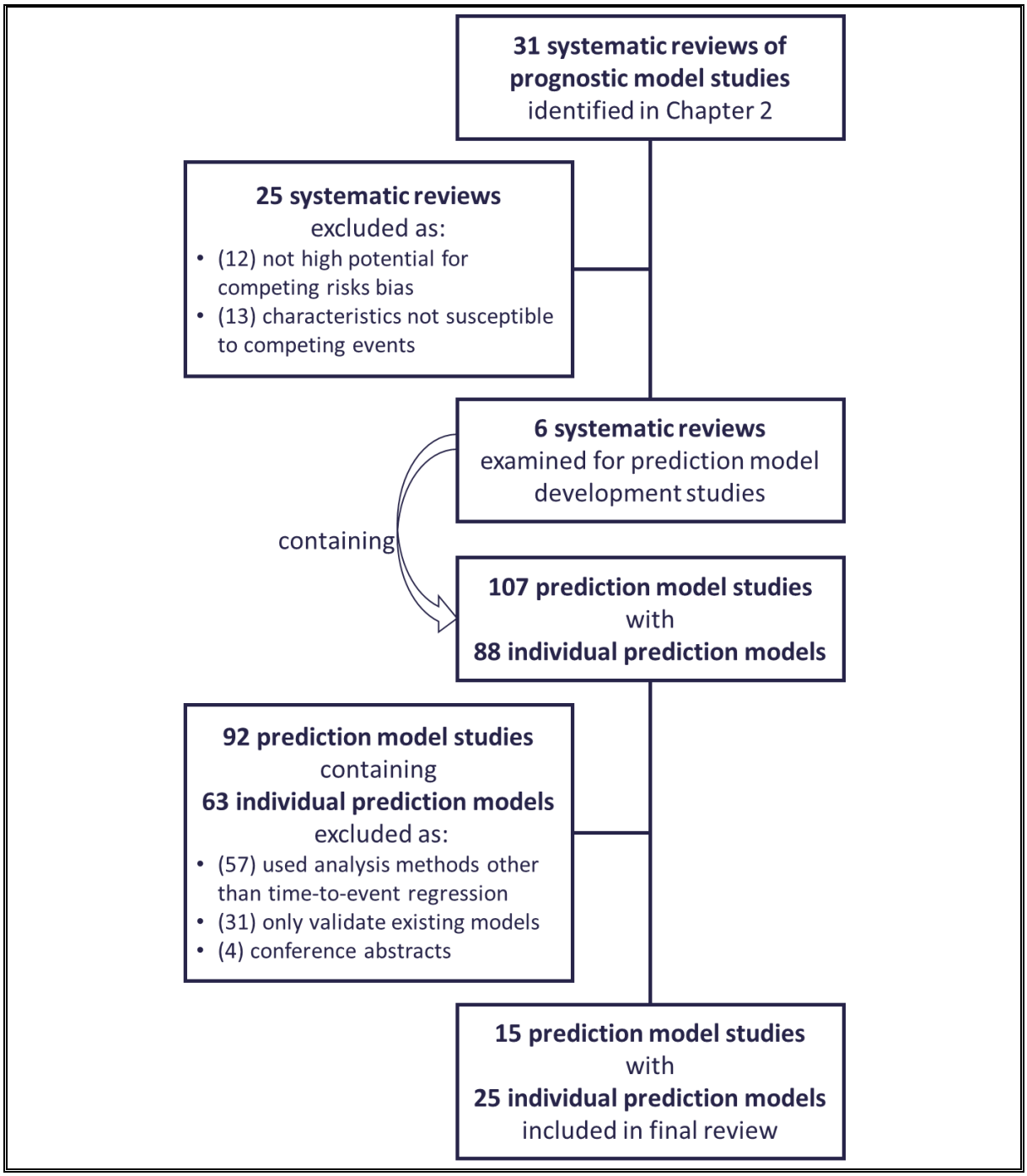
For this review, a narrative synthesis of the information extracted from the prediction model studies was conducted. The preliminary synthesis consisted of tabulation and textual descriptions of the extracted information.

3.3 Results: Review of prediction model development studies

3.3.1 Search and selection of relevant prediction model studies and individual prediction models

A flow diagram depicting the selection process for prediction model development studies included in this review is provided in Figure 3.1.

Figure 3.1: Flow diagram of prediction model study article selection process for this review



The initial selection criteria identified six systematic reviews with high potential for competing risks bias and with population characteristics that were susceptible to competing events. Further information on reasons to include/exclude these systematic reviews is provided in Appendix IV. The six included systematic reviews contained 107 prediction model studies which developed a total of 88 individual prediction models. All 107 published prediction model study articles were screened, and 15 (14.0%) were identified as eligible for this review base on the five criteria listed in Section 3.2.1. Further information on reasons to include/exclude these prediction model studies is provided in Appendix VI. A summary of the 15 included prediction model articles that were identified from the 6 included systematic reviews is provided in Table 3.3. The extracted information for items 1 to 4 are now summarised for the 15 models.

Table 3.3: Prediction model development articles included in this review

Systematic review reference	Total number of prediction model studies	Total number of individual prediction models	Prediction model studies included in this review
(Ayerbe et al., 2016)	12	13	(Sekhri et al., 2008)
(Hilkens et al., 2016)	5	5	(Ariesen et al., 2006) (Cuschieri et al., 2014)
(O'Caoimh et al., 2015)	46	23	(Carey et al., 2008) (Schonberg et al., 2009)
(Salz et al., 2015)	14	14	(Bevilacqua et al., 2012) (Briganti et al., 2010) (Ezaz et al., 2014) (Kovalchik et al., 2012) (Mathieu et al., 2014) (Romond et al., 2012) (Travis et al., 2005)
(Walsh et al., 2016)	12	18	(Nakagawa et al., 2008) (Nyberg and Gustafson, 1997)
(Williams et al., 2016)	18	15	(Hippisley-Cox and Coupland, 2012)

3.3.2 Item 1: What were the characteristics of each prediction model study?

The characteristics of the 15 included prediction model studies are summarised in Appendix VI. In brief, the majority of the prediction model studies (9, 60.0%) develop one individual prediction model; only two studies (13.3%) developed two models, and

four studies (26.7%) developed three models. Most prediction model studies (11, 73.3%) used cohort study data to develop the prediction models. Two (13.3%) prediction model studies used data from nested case-control studies to develop prediction models, one (6.7%) study used only randomised controlled trial (RCT) data, and another (6.7%) used a combination of cohort and RCT data. The total number of participants included in the prediction model studies ranged from 135 to 3,587,653 participants. A number of prediction model studies used bootstrap resampling for model validation (Ariesen et al., 2006, Bevilacqua et al., 2012, Briganti et al., 2010, Cuschieri et al., 2014, Romond et al., 2012). Others opted to split the study participants into development and validation cohorts (Carey et al., 2008, Ezaz et al., 2014, Hippisley-Cox and Coupland, 2012, Schonberg et al., 2009), while one (Kovalchik et al., 2012) sought out additional participants to externally validate the prediction models.

3.3.3 Item 2: What is the potential for competing risks bias affecting each individual prediction model?

The potential for competing risks bias was assessed for each of the 25 individual prediction models identified in the 15 prediction model studies. Information relating to three criteria (prediction model outcome, baseline population, and prediction horizon) were extracted for each individual prediction model and are presented in Table 3.4. Each of the criteria were assessed independently prior to being combined to determine the overall potential for competing risks bias. The results are summarised below.

Table 3.4: The potential for competing risk bias in included individual prediction models developed in the prediction model studies

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcome		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons
		Outcomes of the prediction models	Does outcome contain all-cause mortality?	Disease and health state of prediction model population	Population age at baseline	Prediction horizon of prediction model
(Sekhri et al., 2008)	Resting ECG model	Death due to coronary heart disease or non-fatal acute coronary syndrome	No	Suspected angina	Mean 55 (SD 13) years	6 years
	Summary exercise ECG model				Mean 55 (SD 13) years	
	Detailed exersize ECG model				Mean 54 (SD 11) years	
(Ariesen et al., 2006)		Intracerebral haemorrhage	No	Patients with ischaemic stroke or transient ischemic attack	Mean 64 (SD 10) years*	Up to 5 years¥
(Cuschieri et al., 2014)		Acute gastrointestinal (GI) bleeding	No	Patients with myocardial infarction & prescribed clopidogrel	GI bleed:Mean 66.2 (SD 10.4) years No GI bleed: Mean 62.4 (SD 9.9) years	Up to 8 years¥
(Carey et al., 2008)		All-cause mortality	Yes	Community living frail elderly people with long-term care needs	Mean 79 (SD 9) years	Up to 6 years¥

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcome		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons
		Outcomes of the prediction models	Does outcome contain all-cause mortality?	Disease and health state of prediction model population	Population age at baseline	Prediction horizon of prediction model
(Schonberg et al., 2009)		5-year mortality	Yes	Community dwelling adults aged 65 and older	65-69: 26.6% 70-74: 26.6% 75-79: 21.9% 80-84: 14.6% 85+: 10.4%	5 years
(Bevilacqua et al., 2012)	Preoperative model Within 6 months model 6 months or later model	Lymphedema	No	Axillary lymph node dissection in breast cancer	≤55 years: 582 (55.2%) >55 years: 472 (44.8%)	5 years
(Briganti et al., 2010)		Erectile function recovery	No	Prostate cancer treated with bilateral nerve sparing prostsectomy	Mean: 61.9 years Median: 62 years	3 years
(Ezaz et al., 2014)		Heart failure and cardiomyopathy	No	Patients receiving adjuvant trastuzumab therapy for early-stage breast cancer	Mean 73.6 (SD 5.3) years	3 years

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcome		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons
		Outcomes of the prediction models	Does outcome contain all-cause mortality?	Disease and health state of prediction model population	Population age at baseline	Prediction horizon of prediction model
(Kovalchik et al., 2012)	Self reported risk factor model	Second primary thyroid cancer	No	Survivors of childhood cancers	<5 years: 41% 5-9 years: 22% 10-14 years: 20% 15+ years: 18%	20 years
	Medical record abstraction model					
	All available information model					
(Mathieu et al., 2014)	Urinary toxicity model	Global urinary toxicity grade ≥2	No	Prostate cancer radiotherapy	Mean: 68 years Range: 45 to 83 years	5 years
	Urinary frequency model	Urinary frequency grade ≥2	No			
	Dsuria model	Dysuria grade ≥2	No			
(Romond et al., 2012)		Severe congestive heart failure or cardiac death	No	Patients with node-positive breast cancer	Mean 49 years	5 year
(Travis et al., 2005)	Without counselling model	Breast cancer	No	Young women treated for Hodgkin's lymphoma	Mean: 22 years Median: 22 years	30 years
	With counselling model					
(Nakagawa et al., 2008)		Falls	No	Stroke inpatients in convalescent rehabilitation wards	Mean 69.7 (SD 12.1) years	180 days¥

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcome		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons
		Outcomes of the prediction models	Does outcome contain all-cause mortality?	Disease and health state of prediction model population	Population age at baseline	Prediction horizon of prediction model
(Nyberg and Gustafson, 1997)		Falls	No	Patients in stroke rehabilitation	Mean 74.8 (SD 8.9) years	56 days[¥]
(Hippisley-Cox and Coupland, 2012)	Female model	Colorectal cancer	No	Women with suspected colorectal cancer	Mean 50.1 (SD 15.0) years	2 years
	Male model			Men with suspected colorectal cancer		

***Information gained from appendix or supplementary material.**
[¥]Follow-up measure given as prediction horizon not directly reported in article.
ECG = Electrocardiogram

Criterion for competing risk bias 1: The prediction model investigates outcomes other than all-cause mortality

As mentioned, it is fundamental to consider prediction model outcomes when assessing for the risk of competing risks bias. Individual prediction models which predict the risk of all-cause mortality are unlikely to contain competing events, as few events can prevent the occurrence of death. Information on outcomes predicted by the individual prediction models is displayed in Table 3.4. All-cause mortality was the outcome for two (8.0%) individual prediction models, developed in two prediction model studies (Carey et al., 2008, Schonberg et al., 2009). The remaining 23 (92.0%) individual prediction models, developed in 13 prediction model studies, predicted outcomes other than all-cause mortality, hence met this criterion.

Criterion for competing risk bias 2: The baseline population contains frail and/or elderly populations

The susceptibility of the prediction model study populations is important when assessing the risk of competing risks bias. Competing events are more likely to be present in individual prediction models developed in elderly and frail populations, thus the models have an increased risk of competing risks bias. Summary information of the baseline populations used to develop each individual prediction model is displayed in Table 3.4. All of the individual prediction models were either developed in frail populations, such as those with or surviving cancer (16, 64.0% models from 8, 53.3% prediction model studies), or were developed in populations which included persons over 60 years of age (20, 80% models from 13, 86.7% prediction model studies). Thus, all included prediction models met this criterion.

Criterion for competing risk bias 3: The prediction horizon is sufficiently long to enable competing events to occur

Assessing the duration of the prediction horizon can help to determine the risk of competing risks bias, as a longer prediction horizon enables a greater number of competing events to occur. Thus individual prediction models with long prediction horizons have an increased risk of competing risks bias. Information summarizing the prediction horizons of the individual prediction models is provided in Table 3.4. Two (8.0%) individual prediction models, developed in two (13.3%) prediction model studies, had prediction horizons shorter than a year; these models made predictions at 56 and 180 days (Nakagawa et al., 2008, Nyberg and Gustafson, 1997). The remaining individual prediction models (23, 92.0% models developed in 13, 86.7% prediction model studies) had prediction horizons over 1 year, thus met this criteria. Five (20%) of the individual prediction models, developed in two (13.3%) prediction model studies, made predictions after 20 and 30 years (Kovalchik et al., 2012, Travis et al., 2005).

Overall assessment of the potential for competing risks bias within each individual prediction model

The three criteria, discussed in detail above, were examined and an assessment of the potential for competing risks bias in each individual prediction model was performed. An overview of the final classification of the individual prediction models is presented in Table 3.5.Only two (8.0%) of the individual prediction model studies predicted all-cause mortality outcomes (Criterion 1 not present) (Carey et al., 2008, Schonberg et al., 2009). As it is unlikely that competing events would prevent the occurrence of all-cause mortality, these individual prediction models were categorised as no potential for competing risks bias. The remaining 23 (92.0%) individual prediction models predicted outcomes other than all-cause mortality (Criterion 1 present), and were developed in populations that were considered to be frail and elderly (Criterion 2 present). Two (8.0%) of these individual prediction models, from two (13.3%) prediction model studies (Nakagawa et al., 2008, Nyberg and Gustafson, 1997), had prediction

horizons which were less than 1 year (Criterion 3 not present). These prediction horizons were considered to be sufficiently short to reduce the likelihood of competing events, and thus these prediction models were categorised as moderate potential for competing risk bias. Finally, the majority (21, 84.0% models from 11, 73.3% prediction model studies) of the individual prediction models met all three criteria, and were classified as high potential for competing risks bias.

The 21 (84.0%) individual prediction models classified as high potential for competing risks bias are likely to have been developed in the presence of competing events. Thus, in the sample evaluated, a large majority were susceptible to competing risks bias affecting their prediction model development.

Potential competing events for each prediction model study

A list of potential competing events likely to prevent the prediction model outcomes from occurring is provided in Appendix VII. In brief, potential competing events were identified for all but two (13.3%) of the prediction model studies; as these only contained individual prediction models which predicted all-cause mortality (Carey et al., 2008, Schonberg et al., 2009). Mortality was identified as a potential competing event for the remaining 13 (86.7%) prediction model studies. Further potential competing events included; complete immobility, as immobility significantly alters the risk of falling; removal of the thyroid gland, as this eradicates the risk of recurrent thyroid cancer; and recurrence of the primary cancer or development of a secondary cancer, as these events would meaningfully alter the risk of any future cancer events from occurring.

Table 3.5: An assessment of the potential for competing risk bias in each individual prediction model

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
		Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Carey et al., 2008)		No	All-cause mortality outcome					None
(Schonberg et al., 2009)		No	All-cause mortality outcome					None
(Nakagawa et al., 2008)		Yes	Other outcome	Yes	Stroke indicative of frail population	No	Short prediction horizon makes competng events unlikely	Moderate
(Nyberg and Gustafson, 1997)		Yes	Other outcome	Yes	Stroke indicative of frail population	No	Short prediction horizon makes competing events unlikely	Moderate
(Sekhri et al., 2008)	**Resting ECG model**	Yes	Other outcome	Yes	Angina indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	Summary exercise ECG model	Yes	Other outcome	Yes	Angina indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	Detailed exersize ECG model	Yes	Other outcome	Yes	Angina indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
		Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Ariesen et al., 2006)		Yes	Other outcome	Yes	Stroke indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Cuschieri et al., 2014)		Yes	Other outcome	Yes	Myocardial infarction indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Bevilacqua et al., 2012)	**Preoperative model**	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	Within 6 months model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	6 months or later model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Briganti et al., 2010)		Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
		Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
(Ezaz et al., 2014)		Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Kovalchik et al., 2012)	**Self reported risk factor model**	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	Medical record abstraction model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	All available information model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Mathieu et al., 2014)	**Urinary toxicity model**	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	Urinary frequency model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High

Prediction model study reference	Model	Criterion for competing risk bias 1: Prediction model outcomes		Criterion for competing risk bias 2: Baseline populations		Criterion for competing risk bias 3: Prediction horizons		Overall potential for competing risks bias
		Outcomes other than all-cause mortality?	Justification	Frail or elderly baseline population?	Justification	Prediction horizon at least 1 year?	Justification	
	Dsuria model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Romond et al., 2012)		Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Travis et al., 2005)	**Without counselling model**	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	With counselling model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
(Hippisley-Cox and Coupland, 2012)	**Feamle model**	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High
	Male model	Yes	Other outcome	Yes	Cancer indicative of frail population	Yes	Prediction horizon sufficiently long for competing events to occur	High

3.3.4 Item 3: Were competing events reported in the published prediction model study articles?

Information about the reporting of competing events in the published prediction model study articles are reported by individual prediction models in Table 3.6. The observed number of prediction model events was reported for all but one of the individual prediction models (Carey et al., 2008). However, numbers of observed competing events were not reported for the majority of the individual prediction models (15, 60.0% models from 11, 73.3% prediction model studies); though competing events were not expected to be present for two of these (Carey et al., 2008, Schonberg et al., 2009). The number of competing events was reported (either partially or fully) for ten (40.0%) individual prediction models from four (26.7%) prediction model studies (Bevilacqua et al., 2012, Kovalchik et al., 2012, Nyberg and Gustafson, 1997, Sekhri et al., 2008). The proportion of the total number of events which were competing events was calculated for these individual prediction models; this proportion was found to range from 12.5% to 94.0%. The proportion of all observed events which are competing has been shown to be positively associated with the amount of competing risk bias (van Walraven and Hawken, 2016), thus confirming the importance of appropriately accounting for competing events in prediction model studies.

Kaplan-Meier curves were presented and discussed in 11 (73.3%) prediction model studies, which developed 16 (64.0%) individual prediction models. In the presence of competing events Kaplan-Meier estimates are known to be inflated, thus almost three-quarters of the prediction model studies are likely to have reported inflated absolute risk estimates.

Table 3.6: Reporting of competing events in published prediction model study articles assessed for each individual prediction model

Prediction model study reference	Model	Reported number of events				Key competing risks terms		Prognostic factors	
		Prediction model events	Competing events	Total observed events	Proportion competing events	Terms used	Kaplan-Meier curves	Number in final model	Associated with mortality
(Sekhri et al., 2008)	**Resting ECG model**	576 (7%)	Deaths not reported; 465 (6%) PTCA or CABG	1,041	44.7%	No	No	7	Age, diabetes
	Summary exercise ECG model	351 (7%)	Deaths not reported; 354 (7%) PTCA or CABG	705	50.2%			9	Age, diabetes
	Detailed exercise ECG model	110 (8%)	Deaths not reported; 87 (6%) PTCA or CABG	197	44.2%			9	Age, diabetes
(Ariesen et al., 2006)		107 (1%)	Deaths not reported			No	Yes	4	Age
(Cuschieri et al., 2014)		107 (3%)	Deaths not reported			No	Yes	5	Age, diabetes, chronic liver disease, chronic kidney disease
(Carey et al., 2008)		Not reported				No	Yes	8	Age, malignant neoplasm, congestive heart failure, chronic obstructive pulmonary disease, renal failure

Prediction model study reference	Model	Reported number of events				Key competing risks terms		Prognostic factors	
		Prediction model events	Competing events	Total observed events	Proportion competing events	Terms used	Kaplan-Meier curves	Number in final model	Associated with mortality
(Schonberg et al., 2009)		4,061 (17%)				No	Yes	11	Age, chronic obstructive pulmonary disease, diabetes, cancer
(Bevilacqua et al., 2012)	**Preoperative model**	247 (23%)	171 (16%) deaths	418	40.9%	No [1]	Yes	3	Age
	Within 6 months model	247 (23%)	171 (16%) deaths	418	40.9%			5	Age
	6 months or later model	247 (23%)	171 (16%) deaths	418	40.9%			7	Age
(Briganti et al., 2010)		252 (58%)	Deaths not reported			No	Yes	3	Age, Charlson comorbidity index score
(Ezaz et al., 2014)		155 (19%)	Deaths not reported			No	No	7	Age, coronary artery disease, atrial fibrilation, diabetes, hypertension, renal failure
(Kovalchik et al., 2012)	**Self-reported risk factor model**	159 (1%)	2,483 (20%) competing events[2]	2,642	94.0%	Yes	No	5	Age, Hodgkin lymphoma
	Medical record abstraction model	159 (1%)	2,483 (20%) competing events[2]	2,642	94.0%			7	Age
	All available information model	159 (1%)	2,483 (20%) competing events[2]	2,642	94.0%			6	

Prediction model study reference	Model	Reported number of events				Key competing risks terms		Prognostic factors	
		Prediction model events	Competing events	Total observed events	Proportion competing events	Terms used	Kaplan-Meier curves	Number in final model	Associated with mortality
(Mathieu et al., 2014)	**Urinary toxicity model**	183 (19%)	Deaths not reported			No[1]	Yes	2	
	Urinary frequency model	92 (10%)	Deaths not reported					3	Diabetes
	Dysuria model	36 (4%)	Deaths not reported					2	
(Romond et al., 2012)		37 (4%)	Deaths not reported			Yes	Yes	2	Age
(Travis et al., 2005)	**Without counselling model**	105[3]	Deaths not reported[4]			Yes	No	3	Age
	With counselling model	105[3]	Deaths not reported[4]					4	Age
(Nakagawa et al., 2008)		270 (38%)	Deaths not reported			No	Yes	7	Hasegawa's dementia scale
(Nyberg and Gustafson, 1997)		49 (36%)	Deaths not reported; 7(5%) immobile	56	12.5%	No	Yes	8	
(Hippisley-Cox and Coupland, 2012)	**Female model**	4798 (<1%)[5]	Deaths not reported			No	Yes	7	Age
	Male model		Deaths not reported					9	Age

Prediction model study reference	Model	Reported number of events				Key competing risks terms		Prognostic factors	
		Prediction model events	Competing events	Total observed events	Proportion competing events	Terms used	Kaplan-Meier curves	Number in final model	Associated with mortality

(1) Cumulative incidence calculated using Kaplan-Meier methods
(2) Obtained from supplementary table S3, competing events consist of other secondary primary cancer, thyroid removal, or death.
(3) Nested case-control study so percentage of events not reported as not appropriate.
(4) Nested case-sontrol study reports hazard rates per 100 000 person-years (population based measures) instead.
(5) Number of events reported for prediction model study, not stratified by sex as individual models are.
PTCA = percutaneous transluminal coronary angioplasty; CABG = coronary artery bypass grafting;

Key competing risks terms were reported in three (20.0%) prediction model studies, which developed four (16.0%) individual prediction models (Kovalchik et al., 2012, Romond et al., 2012, Travis et al., 2005). Competing risks terms were used when reporting potential competing events, as illustrated in Box 3.1, and to describe the statistical modelling methods used to develop the individual prediction models, as illustrated in Box 3.2. Other instances where competing risks terms were used are cited in Appendix VIII.

Box 3.1: Competing risks terms reporting potential competing events

"Competing events for [second primary thyroid cancer] were death, self-reported complete removal of the thyroid gland, and other second primary cancers…"
(Kovalchik et al., 2012)

"The competing events were any first event of recurrence, second primary cancers, and deaths precluding [cardiac events]."
(Romond et al., 2012)

"We estimated this future risk, taking into account … competing causes of death."
(Travis et al., 2005)

Box 3.2: Competing risks terms describing statistical modelling methods

"Estimates of absolute risk combined semiparametric estimates of baseline incidences and [relative risks] for [second primary thyroid cancer] and ***competing risks*** *… **competing event** [relative risks] were estimated from the [Childhood Cancer Survivor Study] cohort only...Hazard models for … **competing event** models for [model 1], [model 2], and [model 3] followed a Cox proportional hazards model."*
(Kovalchik et al., 2012)

*"The cumulative proportions of [cardiac events] were estimated and compared by using the **cumulative incidence** function ... The Cox **cause-specific** proportional hazards model was used to evaluate the association between time to [congestive heart failure] and cardiac risks factors. A parametric regression model on **cause specific subdistribution** hazard was used to build a prediction model for 5-year probability of developing [cardiac events] with 95% point-wise CIs, adjusting for significant risk factors."*
(Romond et al., 2012)

*"To compute cumulative absolute risks of breast cancer, we used modified standardized incidence ratios to relate cohort breast cancer risks to those in the general population, enabling application of population-based breast cancer rates, and we allowed for **competing risks** by using population-based mortality rates in female [Hodgkin's lymphoma] survivors."*
(Travis et al., 2005)

3.3.5 Item 4: How were competing events managed in each prediction model study?

Information about the management of competing events in each of the 15 prediction model studies is reported in Table 3.7. Statistical methods (such as competing risks regression) were applied to appropriately manage competing events in six (24.0%) individual prediction models from three (20.0%) prediction model studies (Kovalchik et al., 2012, Romond et al., 2012, Travis et al., 2005). Competing risks bias should not be present as the competing events have been appropriately managed. Competing events were censored during the development of nine (36.0%) individual prediction models, from four (26.7%) prognostic model studies (Bevilacqua et al., 2012, Hippisley-Cox and Coupland, 2012, Nyberg and Gustafson, 1997, Sekhri et al., 2008), and prior to the development of one individual prediction model (Sekhri et al., 2008). Competing risks bias is likely to affect these individual prediction models as censoring or excluding competing events causes inflated absolute risk estimates. The eight (53.3%) remaining prediction model studies, which developed ten (40.0%) individual prediction models, did not report on the management of competing events. While two of these (Carey et al., 2008, Schonberg et al., 2009) were classified as no potential for competing risks bias, the remaining prediction model studies were classified with moderate or high potential for competing risks bias, thus it is unlikely that no competing events occurred in these studies. For example, one cohort study followed 12,648 stroke patients, with a mean age of 64, for 5 years (Ariesen et al., 2006); The study does not report any deaths (competing events) or how these were managed in the study, yet it is unlikely that none occurred.

Table 3.7: Management of competing events in each prediction model study

Prediction model study reference	Potential for competing risks bias	Prediction model outcomes	Competing risks statistical methods used	Competing events censored	Competing events excluded	Validation method
(Sekhri et al., 2008)	High	Death due to coronary heart disease or non-fatal acute coronary syndrome	No	Yes	No	Apparent
(Ariesen et al., 2006)	High	Intracerebral haemorrhage	No	No	No	Internal
(Cuschieri et al., 2014)	High	Acute gastrointestinal (GI) bleeding	No	No	No	Apparent
(Carey et al., 2008)	No	All-cause mortality	No	No	No	Internal
(Schonberg et al., 2009)	No	5-year mortality	No	No	No	Internal
(Bevilacqua et al., 2012)	High	Lymphedema	No	Yes	No	Internal
(Briganti et al., 2010)	High	Erectile function recovery	No	No	No	Internal
(Ezaz et al., 2014)	High	Heart failure and cardiomyopathy	No	No	No	Internal
(Kovalchik et al., 2012)	High	Second primary thyroid cancer	Yes	No	No	External
(Mathieu et al., 2014)	High	Urinary toxicity, urinary frequency, dysuria	No	No	No	Apparent
(Romond et al., 2012)	High	Severe congestive heart failure or cardiac death	Yes	No	No	Internal
(Travis et al., 2005)	High	Breast cancer	Yes	No	No	None
(Nakagawa et al., 2008)	Moderate	Falls	No	No	No	Apparent
(Nyberg and Gustafson, 1997)	Moderate	Falls	No	Yes (deaths)	Yes (immobility)	Apparent

Prediction model study reference	Potential for competing risks bias	Prediction model outcomes	Competing risks statistical methods used	Competing events censored	Competing events excluded	Validation method
(Hippisley-Cox and Coupland, 2012)	High	Colorectal cancer	No	Yes	No	Internal

The predictive performance of each developed model was evaluated in some way for the majority of models (23, 92.0% models from 14, 93.3% prediction model studies). Apparent performance measures were reported for nine (36.0%) individual prediction models from five (33.3%) prediction model studies; internal validation methods were reported for 11 (44.0%) prediction models from eight (53.3%) studies; and external validation was reported for three (12.0%) models from one (6.7%) study. None of the prediction model studies reported on how competing events were managed during model validation.

3.4 Discussion

In this chapter, a review of the presence, reporting, and management of competing events in prediction model development studies was conducted. In total, 25 individual prediction models were evaluated, as identified from 15 prediction model studies likely to be affected by competing events. The key findings and conclusions of this review are summarised in Box 3.3 and discussed below.

Box 3.3: Key findings

1. Mortality is a common competing event for prediction model outcomes.
2. Prediction model development articles rarely report competing risks, even when studies are conducted in clinical settings where competing events are a likely issue.
3. The management of competing events in prediction model development studies is inconsistent and often not appropriate.

3.4.1 Key findings

The key findings of this review are presented and discussed in more detail:

3.4.1.1 Mortality is a common competing event in prediction models.

Of the 25 individual prediction models included in this review, only 8.0% predicted all-cause mortality, and a further 16.0% predicted cause-specific mortality. Thus mortality, either all-cause or other-cause, was considered a competing event for 92.0% of the prediction models identified in this review. Further, all of the individual prediction models were developed in either frail or elderly populations, and these populations are particularly susceptible to the competing risk of death (Berry et al., 2010). More needs to be done to highlight the importance of appropriately accounting for mortality as a competing event in prediction model studies in elderly or frail populations.

3.4.1.2 Prediction model development articles rarely report competing risks, even when studies are conducted in clinical settings where competing events are a likely issue.

The included prediction model studies were identified for having high potential for competing risks bias and population characteristics that were susceptible to competing events. Despite this, competing events were poorly reported in the published prediction model articles. Key terms related to competing events were not mentioned in 80% of the prediction model study articles (which developed 84.0% of individual prediction models). Further, the number of competing events was not reported in 73.3% of the prediction model study articles (60.0% of individual prediction models). It is unlikely that competing events were not present in the majority of the studies, given the selection criteria and potential for competing risks bias. Thus, the lack of reporting of competing events may indicate the lack of awareness of competing risks bias as an issue in prediction model research.

3.4.1.3 The management of competing events in prediction model development studies is inconsistent and often not appropriate.

The management of competing events during the development of the prediction models was inconsistent; for example, statistical methods to appropriately handle competing events were applied in only 20.0% of the prediction model studies (24.0% of individual prediction models), whilst competing events were censored in 26.7% of the prediction model studies (36.0% of individual prediction models). The management of competing events in the validation of prediction models was not discussed in any of the articles. Not appropriately managing competing events when they are present can lead to bias, inflated absolute risk predictions. Moreover, not accounting for the competing events when validating prediction models can lead to bias in predictive performance measures. Therefore, competing events need to be handled more

appropriately in new prediction model development studies where competing events are an issue.

3.4.2 Limitations and further research

This work adds to existing reviews and empirical evaluations of whether competing risks have been accounted for in medical research. Previous systematic reviews investigating competing risk biases have focused on articles containing Kaplan-Meier curves (Schumacher et al., 2016, Walraven and McAlister, 2016), studies with population susceptible to competing risks (Koller et al., 2012), or published randomised controlled trials (Austin and Fine, 2017). This systematic review focuses on competing risks bias in prediction model development studies and thus provides a novel insight into the presence, reporting, and management of competing events during the prediction model development process. The review highlights the high presence of death as a competing event in prediction model studies in elderly and frail populations. Nevertheless, the review does not encompass all prediction model development research. The high presence of competing events is unlikely to be generalizable to all prediction model development studies, given the selection of articles from of high potential for competing risks bias systematic reviews. However, focusing on studies likely to be affected by competing events allowed a concentrated investigation into the reporting and management of competing events in model development articles where competing events were likely to be present and cause bias results if inappropriately accounted for.

There has been relatively little research about the impact of competing events in relation to prediction model research. The way in which competing events may affect prediction model outcomes (if not appropriately accounted for) include: erroneous estimates of prognostic factor associations, inflated absolute risk predictions, mis-calibration, and inaccurate risk group allocation. This chapter has highlighted the high

proportion of prediction model development studies which do not appropriately account for competing events. However, further research is needed to evaluate the impact of this bias on the prediction model outcomes (such as absolute risk predictions, and measures of calibration and discrimination). In the following chapters the statistical methods available to appropriately account for competing events will be discussed and applied to develop a prediction model for the risk of antenatal adverse events using data from the Prediction of Risks in Early onset Pre-eclampsia (PREP) study (Thangaratinam et al., 2017).

4 A COMPARISON OF TIME-TO-EVENT PROGNOSTIC MODELS DEVELOPED USING COX AND FLEXIBLE PARAMETRIC METHODS

4.1 Introduction

Before investigating the application and impact of competing risks methods, it is helpful to introduce and apply standard time-to-event methods to develop prognostic models. Thus, in this chapter two prognostic models are developed using different popular time-to-event methods, namely Cox proportional hazards regression (Cox, 1972a) and Royston-Parmar flexible parametric regression (Royston and Parmar, 2002). Cox proportional hazards regression is commonly applied to develop prognostic models for time-to-event outcomes (Collins et al., 2015, Royston et al., 2009). However, due to its semi-parametric form, Cox regression does not estimate the baseline hazard function and consequently does not allow for direct estimation of absolute risks (the motivation for prognostic model research). Parametric time-to-event models directly estimate the baseline hazard function and thus provide estimates of absolute risks over time, making them appropriate for use in prognostic model research (Snell, 2015). Some parametric models make strong assumptions about the shape of the hazard function which are simplistic and unlikely, such as exponential models which assume the hazard function is constant over time. Royston-Parmar flexible parametric regression estimates the baseline hazard function using restricted cubic splines, which are capable of capturing complex hazard functions (Royston and Parmar, 2002). In this chapter both the Cox and Royston-Parmar methods will be applied to existing study data to develop and internally validate two new prognostic models which predict the risk of antenatal adverse events in women diagnosed with

early onset pre-eclampsia. The resulting models will be compared to illustrate their differences, and the findings will influence the choice of models used in later chapters of this thesis, where extensions to competing risks are considered.

4.1.1 Background: The PREP study

Pre-eclampsia is a serious disorder in pregnancy, characterised by elevated blood pressure and excessive protein in the urine (Sibai, 2003). When this occurs early in the pregnancy (before 34 weeks gestation) it can cause serious complications, several of which may be life-threatening for both the mother and the unborn baby (Sibai, 2003). The varying prognoses of early-onset pre-eclampsia necessitates accurate prediction of complications to allow the timely recognition, referral and treatment of women with a high-risk of complications (Wilkinson, 2011).

Prior to this thesis, two multivariable prognostic models were developed in the Prediction of Risks in Early onset Pre-eclampsia (PREP) study (Thangaratinam et al., 2017). This study utilized data from a prospective cohort of 946 women diagnosed with confirmed early-onset pre-eclampsia between December 2011 and April 2014 from 53 maternity units in the UK. These prognostic models provide predictions for individual risks of adverse maternal outcomes, including delivery of a preterm infant, for women with early-onset pre-eclampsia. One model predicts events prior to discharge, the “PREP-L” logistic regression model, the other predicts events at various time points prior to 34 weeks gestation, the “PREP-S” Royston-Parmar parametric model. Both models were externally validated in the Pre-eclampsia Integrated Estimate of RiSk for mothers (PIERS) cohort (von Dadelszen et al., 2009) and the Pre-eclampsia Eclampsia Trial Amsterdam (PETRA) cohort (Ganzevoort et al., 2005), comprising of a total of 850 women. Further details of the PREP study can be found in (Thangaratinam et al., 2017).

The inclusion of preterm delivery in the composite outcome predicted by both PREP prognostic models was determined through the consensus of an expert panel (Thangaratinam et al., 2017). Delivery of the baby is the only known cure for pre-eclampsia, yet delivery before 34 weeks gestation (preterm delivery) increases the baby's risk of death and neurological disability (Sibai, 2003). Despite this risk, preterm delivery is often offered to prevent further maternal complications and is thus indicative of patients at considerable risk of adverse maternal outcomes (Thangaratinam et al., 2017). As preterm delivery is thus a treatment for high risk individuals, not accounting for preterm delivery in the models may have introduced bias through the treatment paradox.

Treatment paradox bias occurs when the presence of a strong predictor of an adverse event triggers an effective treatment, thereby preventing the occurrence of a number of adverse outcomes (Cheong-See et al., 2016). Had the treatment not been given, a subsequent adverse outcome would have been highly likely to occur; thus, ignoring the treatment weakens the perceived associations between true predictors and the outcome. Hence the expert panel agreed to incorporate preterm delivery into the composite outcome.

Though the use of a composite outcome reduces the risk of treatment paradox bias, combining the outcomes presents a number of limitations. Preterm delivery was the most frequent outcome, thus prognostic model predictions were found to be heavily influenced by the risk of preterm delivery. The composite outcome limits the interpretation of associations between prognostic factors and individual outcomes, as associations with specific outcomes were not investigated. Finally, as preterm delivery, by definition, only occurs prior to 34 weeks gestation, the prediction horizon of the PREP-S model was restricted to 34 weeks. Any events which occurred after this time were censored, thus the PREP-S model is not able to make predictions over a full pregnancy term (40 weeks).

4.1.2 Prediction of antenatal adverse events

To address some of the limitations of the PREP model composite outcome, an alternative outcome, antenatal adverse events, is considered and used throughout this thesis. Antenatal adverse events refer to adverse events which occur *during pregnancy*. Following delivery of the baby (be it preterm or otherwise) the participant is no longer pregnant, and thus no longer able to experience an antenatal adverse event. In this chapter, participants who deliver will be censored at the time of delivery (deliveries will be managed differently in later chapters when competing risks models are utilised). This definition allows established time-to-event regression methods to be applied to develop prognostic models to predict the risks of antenatal adverse events. This alternative outcome definition does not invalidate the original PREP models, which focus on predicting the global risks of adverse outcomes by discharge or 34 weeks. In particular, the models developed in this chapter are not directly comparable to the original PREP models, but may be used alongside the PREP models to further assist in the understanding of the risks to the participants.

4.1.3 Aims

Data from the PREP study cohort will be used to develop and internally validate two new prognostic models to predict the risks of antenatal adverse events. The first will be developed using the Cox proportional hazards method, and the second using the Royston-Parmar flexible parametric method. The parameter estimates of the prediction models will be compared, as will the resulting model predictions, measures of predictive performance, and optimism.

4.2 Methods: Developing two prognostic models

Methods for development and internal validation of the two prognostic models are now discussed.

4.2.1 Outcome definition

The primary outcome, antenatal adverse events, is a composite outcome made up of the events listed in Table 4.1. Patients were considered to be at risk from the date of pre-eclampsia diagnosis. The end of follow-up was defined as the time of occurrence of the first antenatal adverse event or delivery of the baby. Antenatal adverse events by definition exclude postpartum haemorrhages and events which occur following delivery, both of which were included in the original PREP models. Participants were censored at the time of delivery, if this occurred prior to an antenatal adverse event. If both an adverse event and delivery occurred on the same date the outcome was recorded as an antenatal adverse event. A total of 75 antenatal adverse events were observed during the PREP study follow-up (Table 4.1).

Table 4.1: Individual components of antenatal adverse events

Antenatal adverse event	Number (%)
Placental abruption	25 (2.6%)
Transfusion of blood	23 (2.4%)
Eclamptic seizure	11 (1.2%)
Hepatic dysfunction	5 (0.5%)
Pulmonary oedema	4 (0.4%)
Intubation	3 (0.3%)
Acute renal insufficiency	2 (0.2%)
At least 50% FIO2 for > 1hour	1 (0.1%)
Glasgow Coma score <13	1 (0.1%)
Total antenatal adverse events	75 (7.9%)

4.2.2 Candidate prognostic factors

To reduce the risk of overfitting during model development, the number of candidate prognostic factors considered for inclusion in the new models was reduced. The candidate prognostic factors were restricted to those included within the original PREP models that were also available in the PIERS validation study[i]. Abnormal pulse oximetry was not considered as a candidate prognostic factor due to the rarity of its occurrence, as only four cases were observed. The restricted set of candidate prognostic factors is listed in Table 4.2. Maternal age and gestational age are considered to be clinically important risk factors (Thangaratinam et al., 2017), thus were forced into the prognostic models to ensure clinical acceptability. Antihypertensive and magnesium sulphate treatment variables were also forcibly retained to circumvent suboptimal prognostic performance when applied to patients not receiving treatments (Groenwold et al., 2016).

Table 4.2: Candidate prognostic factors for new prognostic models

Maternal characteristics: Maternal age at diagnosis (years), Gestational age at diagnosis (weeks).
Medical history: Count of pre-existing conditions (0, 1, 2 or more) from prespecified list: pre-existing hypertension, renal disease, diabetes mellitus, previous history of pre-eclampsia.
Bedside examination and laboratory tests: Systolic blood pressure (mmHg, highest measurement over 6 hrs), Platelet count (x 109/L), Alanine amino transaminase (IU/l), Serum creatinine µmol/L.
Management at baseline (before or within 1 day of diagnosis): Administration of oral and/or parenteral anti-hypertensives, Administration of magnesium sulphate.

4.2.3 Descriptive analysis of PREP participants

The study participants' follow-up time was investigated, and the median and interquartile range reported. These estimates were calculated using the reverse Kaplan-Meier estimator (Schemper and Smith, 1996), which is calculated in the same

[i]Time-to-event information was not collected during the PETRA study, thus only data from the PIERS study is suitable for external validation of the prognostic models developed in this thesis.

way as the Kaplan-Meier estimate, after the event indicator is reversed so the event of interest is censored and the censored observations become events.

The descriptive analysis of candidate prognostic factors included reporting the frequency and percent of participants in each category of the binary and categorical candidate prognostic factors. Histograms of the distribution of each continuous prognostic factor were inspected to determine which were normally distributed. The mean and standard deviation were reported for factors considered to be normal. The median and inter-quartile range were reported for those which were not considered normal. Finally, correlations between the continuous candidate prognostic factors were examined, none were found to be both strongly (Spearman's rank-order correlation coefficient $|\rho| \geq 0.7$) and significantly ($p < 0.05$) correlated with each other.

4.2.4 Missing information and multiple imputation

The frequency and percent of missing information was reported for each of the candidate prognostic factors. A large proportion of participants (829, 87.5%) had complete data for all outcome and candidate prognostic factor variables. Multiple imputation using chained equations (Buuren and Oudshoorn, 2000) techniques were applied to account for the missing information and avoid a loss in study efficiency and power. Under a missing at random assumption, a total of 20 imputed datasets were created using the *mi impute chained* package in Stata 12. The chained equations included the candidate prognostic factors listed in Table 4.2 (with medical history separated into its individual components), as well as outcome information including a binary variable indicating whether the participant experienced an antenatal adverse event or not, and an estimate of the cumulative hazard function over time (White and Royston, 2009). Missing values from continuous prognostic factors were imputed using the predictive mean matching method (Little, 1988), and those from binary prognostic factors were imputed using augmented logistic regression (White et al., 2010). The

imputed values were visually inspected to determine any outliers. Information on outcome measures was complete, thus outcomes were not imputed.

4.2.5 Time-to-event regression models

Two new prognostic models for predicting the risk of antenatal adverse events in women diagnosed with early onset pre-eclampsia were developed using the PREP study data and different time-to-event methods. The first using Cox proportional hazards regression, the second using Royston-Parmar flexible parametric regression. Details of both of these regression methods are provided in Chapter 1.

Briefly here, Cox proportional hazards regression (Cox, 1972a) was used to develop one prognostic model using the *stcox* estimation command in Stata. This method estimates regression coefficients through maximisation of the partial likelihood function on the hazards scale, as described in Chapter 1. The Efron method (Efron, 1977) was implemented to handle tied failure times, in which the tied failures which occur just before time t are included into the risk set at time t with a probability equal to 1/number of tied events. Breslow's estimate (Equation 1.10) of the baseline cumulative hazard was acquired to estimate the baseline survival function (Breslow, 1972), allowing the calculation of absolute risk estimates over time.

The Royston-Parmar flexible parametric model (Royston and Parmar, 2002) was used to develop the alternative prognostic model using the *stpm2* estimation command in Stata. Restricted cubic splines of log-time are utilised to approximate the log cumulative baseline hazard function. The model regression coefficients and parameters of the spline function are simultaneously estimated through maximum likelihood estimation. The number of knots required to capture the shape of the hazard function was determined by visual inspection of plots of the baseline hazard function for null models (containing no prognostic factors) with varying degrees of freedom (between 2 and 6). A spline with M interior knots is estimated with $M + 1$ degrees of

freedom. Knots were positioned at equally spaced centile values of the distribution of uncensored failure times.

Regardless of the regression method used for development, the form of the prognostic model equation can be written as the cumulative risk of an event occurring prior to time t, as:

$$F_i(t) = 1 - S_0(t)^{\exp(\boldsymbol{\beta}^T \mathbf{X_i})} \qquad \textit{Equation 4.1}$$

Where the linear predictor $\boldsymbol{\beta}^T \mathbf{X_i}$ is the combination of estimated regression coefficients and prognostic factors in the model for the i^{th} participant.

4.2.6 Fractional polynomial terms in multiply imputed data

In order to perform variable selection in multiply imputed data, a stacked approach (Wood et al., 2008) was utilised. This entails stacking the μ imputed datasets to produce one large dataset, and fitting weighted regression models to obtain model parameter estimates and adjusted standard errors. The weights assigned to each observation are calculated as follows:

$$w_j = (1 - f_j)/\mu \qquad \textit{Equation 4.2}$$

In which f_j is the fraction of missing data from the j^{th} prognostic factor. The stacked approach has many advantages over the well-established application of Rubin's rules and inferential framework (Rubin, 2004), including being less computationally demanding and allowing direct estimation of likelihood ratio test statistics. An additional benefit is the extension of the approach to incorporate fractional polynomial terms (Morris et al., 2015), which allows investigations of non-linear associations between continuous prognostic factors and the outcome. Fractional polynomial functions with one dimension (FP1) were investigated for all continuous candidate prognostic factors. The best-fitting FP1 function was determined by estimating the best-fitting exponent

term $\hat{e}$, chosen from the set $e \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$[i], using maximum likelihood-ratio type tests in the stacked data. Centring was permitted for continuous prognostic factors to allow for a more meaningful interpretation of relative risk estimates.

4.2.7 Multivariable analysis and prognostic factor selection

For both the Cox and flexible parametric models, a multivariable analysis was performed to determine the effects of the prognostic factors on the outcome of antenatal adverse events, adjusting for the effect of other included prognostic factors. To determine which of the candidate prognostic factors would be included in the final multivariable prognostic models, a backwards elimination procedure adapted to multiply imputed data and which incorporated fractional polynomial terms (Morris et al., 2015) was applied using the *mfpmi* command in Stata. The procedure is described in Box 4.1.

Box 4.1: Backward elimination procedure adapted for multiply imputed data to incorporate fractional polynomial terms.

Using a stacked dataset, begin with a full model containing all candidate prognostic factors with all continuous prognostic factors included as linear terms.

(1) Using the stacked approach, perform weighted likelihood-ratio tests on the prognostic factors in the model. Order from most to least significant.
(2) By order of significance, working through each prognostic factor in the model. Run the appropriate selection procedure:
Categorical prognostic factors:
Test for removal using weighted likelihood-ratio tests.
If $p < 0.15$ retain in the model, else drop from the model[ii].
Move on to the next significant prognostic factor.
Continuous prognostic factors:
(i) Select the best-fitting FP1 function (as described in 4.2.6).
(ii) Test the model containing the best-fitting FP1 function against the model with the linear term using weighted likelihood-ratio tests.
If $p < 0.15$ retain the FP1 function in the model.
Else test for removal of the linear term using weighted likelihood-ratio tests.
If $p < 0.15$ retain in the model, else drop from the model[ii].
(iii) Move on to the next significant prognostic factor.
(3) Repeat steps 1 and 2 until the model is stable with respect to the retained prognostic factors and their fractional polynomial functional forms.

[i] Conventionally the exponent term of $p = 0$ represents a log-transformation
[ii] Unless prognostic factor is antihypertensive treatment or treatment with magnesium sulphate, which are being forcibly retained in the prognostic models (see 4.2.2).

Maternal and gestational age are retained in the models regardless of their statistical significance to ensure clinical credibility of the final models in line with the original PREP study (Thangaratinam et al., 2017). The treatment variables (antihypertensive and magnesium sulphate treatment) are likewise retained to reduce the risk of suboptimal prognostic performance (Groenwold et al., 2016). A significance threshold of $P < 0.15$ was selected to improve the chances of retaining meaningful prognostic factors in the model. The resulting hazard ratios and 95% confidence intervals from the multivariable models were compared across the two modelling approaches.

4.2.8 Baseline cumulative risk estimates

Estimates of the baseline survival function were calculated using transformations of baseline cumulative hazard functions (Equation 1.7), which were obtained by fitting regression models with the linear predictor from the multivariable prognostic models as an offset term, in each of the imputed datasets separately. Baseline cumulative hazard estimates for the Cox proportional hazards model were obtained using the non-parametric Breslow's estimate (Breslow, 1972) at given time points (Equation 1.10). Baseline log cumulative hazard estimates for the Royston-Parmar flexible parametric model were calculated using the spline functions estimated in the multivariable model (Royston and Lambert, 2011).

An individual's prognostic factor information is combined with the estimated regression coefficients to obtain an individual linear predictor value, $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X_i}$. This is then incorporated into the regression equation alongside the estimated baseline survival function, $S_0(t)$, to obtain individual risk predictions (Equation 4.1).

4.2.9 Sensitivity analysis

A sensitivity analysis was performed to investigate the assumptions made during model development, namely the proportional hazards assumption and independent effects. Both the Cox and Royston-Parmar models, when developed in the way

described above, assume the effects of each prognostic factor on the outcome are constant over time. This assumption was tested by incorporating interactions between the retained prognostic factors and the natural logarithm (ln) of time into the fitted multivariable models. Additionally, interactions between retained prognostic factors were investigated to test the assumption that the effects of each prognostic factor on the outcome were independent. Any statistically significant $(p < 0.05)$ deviations from the assumptions were reported.

4.2.10 Apparent prognostic performance of fitted prognostic models

The apparent performance of the fitted models was assessed in the development data using measures of calibration (overall calibration and calibration slope) and discrimination (Harrell's C-index and R^2_D), as outlined in Section 1.6.3. The model's predictive performance was evaluated overall and at two days, one week, and four weeks. The prognostic performance was assessed in each imputed dataset separately, the imputation-specific measures were then pooled using Rubin's rules (Rubin, 2004).

4.2.11 Internal validation and optimism adjustment

The optimism in apparent performance of the fitted models was assessed through internal validation in 100 bootstrap samples of the original PREP study data. The bootstrap validation procedure repeats the modelling process described above in each bootstrap sample; the procedure is described in Box 4.2.

Box 4.2: Bootstrap procedure for internal validation of fitted prognostic models where there is missing data in the development dataset

(1) Generate a new bootstrap sample containing an equal number of participants as the PREP study by sampling with replacement from the PREP study data.
(2) In the bootstrap sample, develop a prognostic model using the model development process, including:
    (i) Application of multiple imputation methods to create 20 imputed datasets.
    (ii) Performing prognostic factor selection using the backwards elimination procedure outlined in Box 4.1.
    (iii) Calculation of baseline survival estimates appropriate to the regression method being applied.
(3) For each predictive performance statistic (e.g. c-index, calibration slope), assess the average (across imputation datasets) performance of the new models using:
    (i) The bootstrap sample imputation datasets in which the model was developed, referred to as the bootstrap apparent performance.
    (ii) The original PREP study imputed datasets, referred to as the test performance.
(4) For each statistic, calculate the optimism as the difference between the bootstrap apparent performance and the test performance.
(5) Repeat steps 1-4 100 times to obtain an estimate of the average optimism across bootstrap samples.

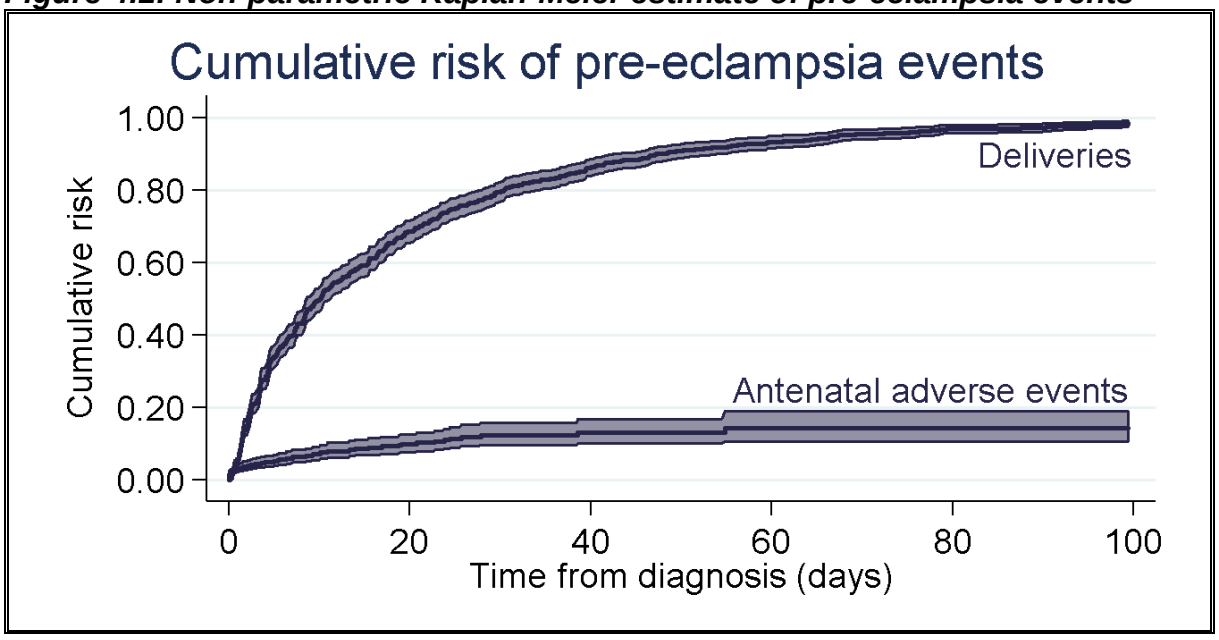
Optimism adjusted prognostic performance measures were calculated by subtracting the average optimism from the apparent prognostic performance measures estimated for the fitted prognostic models previously (see Section 4.2.10). The fitted models were adjusted for optimism by multiplying the fitted regression coefficients by the optimism adjusted calibration slope obtained from the bootstrap procedure. Thus, the optimism adjusted calibration slope represents a uniform shrinkage factor to penalise for overfitting. After shrinkage, and holding the shrunken regression coefficients as known, the baseline survival functions were re-estimated to maintain overall calibration between expected and observed risks.

This process led to two final prediction models, one developed using Cox regression and one using Royston-Parmar models. The final model's included prognostic factors, internal validation model performance, and final optimism adjusted equations were compared.

4.3 Results: Comparison of prognostic models

A total of 947 women with a confirmed diagnosis of pre-eclampsia were followed up for a median of 10.1 days (IQR: 3.4 to 25.3). The longest observed follow-up period for a participant was 126.6 days, none of the participants were lost to follow-up. Antenatal adverse events were observed in 75 (7.9%) women, the remaining 872 patients were followed until delivery of the baby. Non-parametric estimates of cumulative risk of an antenatal adverse event and delivery without an adverse event are depicted using 1-Kaplan-Meier curves in Figure 4.1. A large proportion of participants delivered soon after pre-eclampsia diagnosis; the median time to delivery is 10 days.

Figure 4.1: Non-parametric Kaplan-Meier estimate of pre-eclampsia events



4.3.1 Descriptive analysis of PREP participants

Results of the descriptive analysis of candidate prognostic factors in PREP study participants are given in Table 4.3. The mean maternal age of participants at pre-eclampsia diagnosis was 30.2 (SD 6.1) years, with a median corresponding gestational age of 31.4 weeks (IQR: 28.7 to 32.7). Just under two-thirds of patients did not have any historical medical conditions (63.9%) and around one in ten had two or more

(10.1%). The majority of participants were receiving magnesium sulphate (79.4%), with a minority treated for hypertension (15.2%).

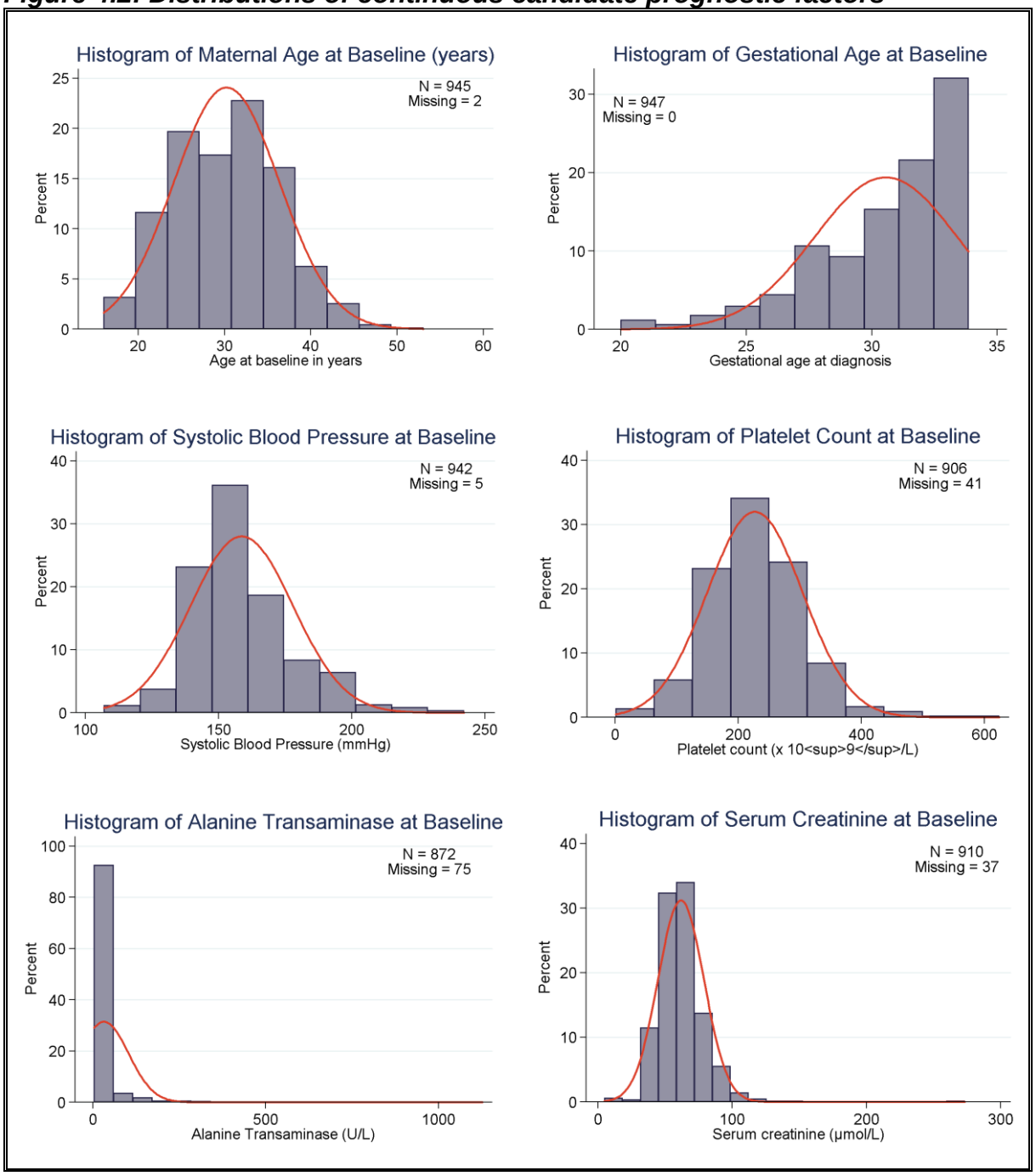
Table 4.3: Descriptive analysis of candidate prognostic factors

Prognostic factor		Summary statistics Mean(SD), N(%), Med [IQR]	Missing N (%)
Maternal age (years)		30.2 (6.1)	2 (0.2%)
Gestational age at diagnosis (weeks)*		31.4 [28.7 to 32.7]	0 (0%)
Medical History¥	**0**	594 (63.9%)	17 (1.8%)
	1	242 (26.0%)	
	2 or more	94 (10.1%)	
Systolic blood pressure		158.6 (19.2)	5 (0.5%)
Platelet count		227.1 (77.7)	41 (4.3%)
Alanine amino transaminase*		17 [13 to 26.5]	75 (7.9%)
Serum Creatinine		61.8 (17.1)	37 (3.9%)
Antihypertensive treatment		751 (79.4%)	1 (0.1%)
Magnesium sulphate treatment		144 (15.2%)	1 (0.1%)

***Median and inter-quartile range (IQR) presented for non-normally distributed factors.**
¥Medical history is a count of the following conditions: chronic hypertension, renal disease, diabetes mellitus, and previous pre-eclampsia.

The continuous prognostic factors of maternal age, systolic blood pressure (SBP), platelet count, and serum creatinine, appeared to be approximately normally distributed, whereas gestational age and alanine transaminase (ALT) were not (Figure 4.2). No strong and significant correlations were found between any of the candidate prognostic factors.

Figure 4.2: Distributions of continuous candidate prognostic factors



One in eight women (12.5%) in the study had missing information for at least one candidate prognostic factor. The candidate prognostic factors with the greatest amount of missing information were the three laboratory tests: ALT (7.9%), platelet count (4.3%), and serum creatinine (3.9%). Multiple imputation using chained equations was applied to create 20 imputed datasets. Imputed values for each candidate prognostic factor were visually inspected and no anomalies were found.

4.3.2 Flexible parametric restricted cubic splines

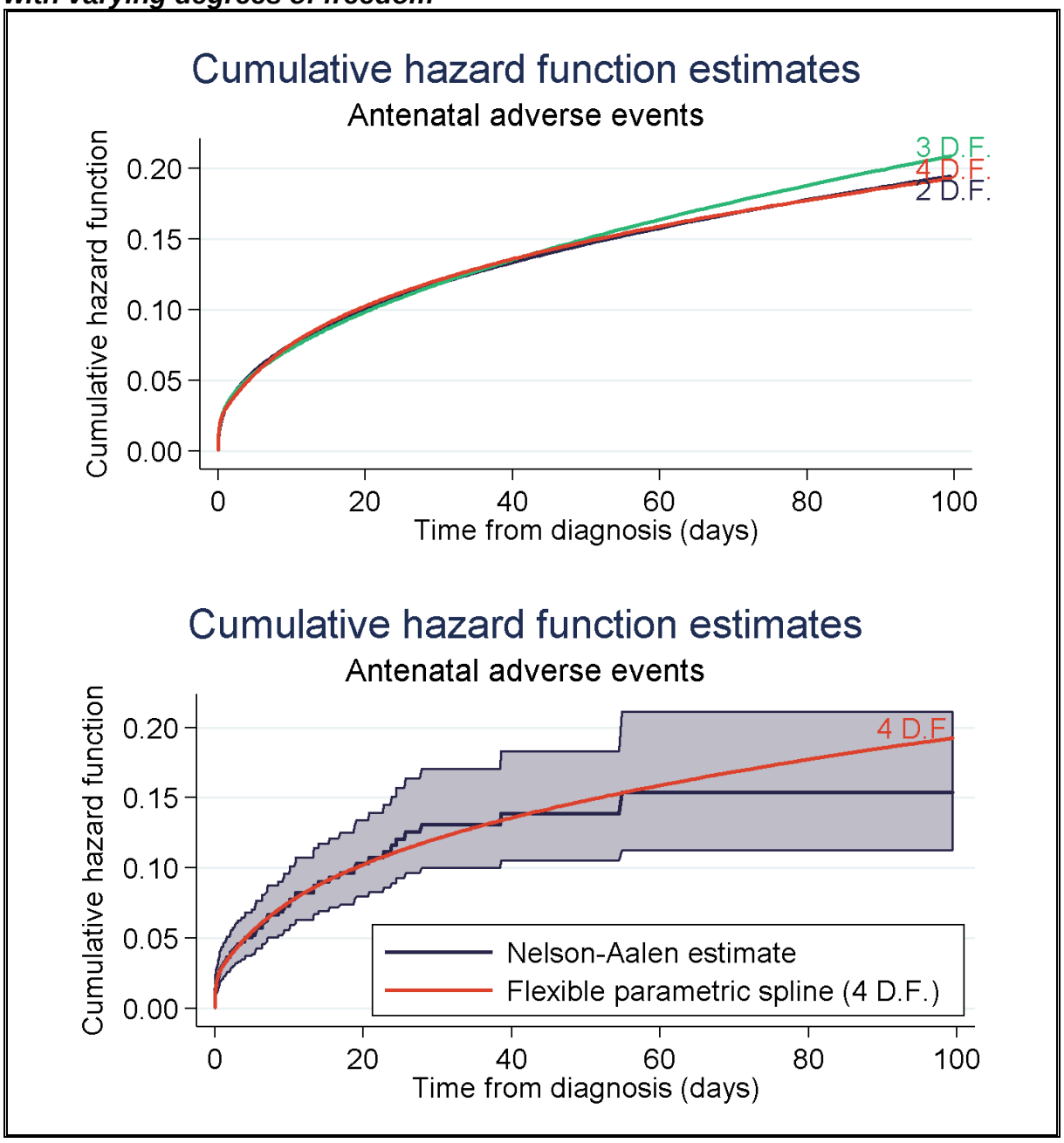
Null Royston-Parmar models, with between one and five internal knots, were fitted to the original PREP study data. The resulting knot locations and, AIC and BIC values of the models are listed in Table 4.4.

Table 4.4: Knot selection for Royston-Parmar flexible parametric model

Number of internal knots	Degrees of freedom	Knot locations (time in days)	AIC	BIC
1	**2**	2.7	801.1	815.6
2	**3**	0.7, 7.0	802.2	821.6
3	**4**	0.4, 2.7, 10.1	802.6	826.9
4	**5**	0.1, 1.3, 5.4, 13.3	803.2	832.3
5	**6**	0.1, 0.7, 2.7, 7.0, 15.4	803.0	837.0

Surprisingly the model with only one internal knot returned the lowest AIC and BIC values (AIC=801.1 and BIC = 815.6), although the differences between AIC and BIC values over the varying number of knots were small. A graphical display of the spline estimates with between two and four degrees of freedom is depicted in Figure 4.3. There is little difference in the cumulative hazard function estimates between the models. The spline functions were compared to a non-parametric Nelson-Aalen cumulative hazard estimate to assess fit. The model with four degrees of freedom (three internal knots) appeared to sufficiently capture the shape of the observed cumulative hazard function (Figure 4.3) so was considered suitable for the remainder of the analysis.

Figure 4.3: Cumulative hazard function estimates using restricted cubic splines with varying degrees of freedom



4.3.3 Multivariable analysis and prognostic factor selection process

The backwards elimination procedure was applied using both regression modelling methods. The resulting hazard ratio estimates and 95% confidence intervals for both models are reported in Table 4.5.

Table 4.5: Multivariable estimates for Cox and Royston-Parmar models

		Cox proportional hazards model		Royston-Parmar flexible parametric model	
		Transformation of prognostic factor (X)	**HR (95% CI)** ***p-value***	**Transformation of prognostic factor (X)**	**HR (95% CI)** ***p-value***
Maternal age (years)		X-30.244	0.975 (0.938, 1.013) *0.192*	X-30.244	0.962 (0.936, 1.011) *0.167*
Gestational age (weeks)		X-30.562	1.018 (0.939, 1.104) *0.665*	X-30.562	1.015 (0.938, 1.098) *0.710*
Medical history[¥]	**1** **2 or more**		0.457 (0.239, 0.871) 0.658 (0.285, 1.518) *0.038*		0.460 (0.242, 0.875) 0.634 (0.275, 1.464) *0.037*
Systolic blood pressure		X-158.634	1.015 (1.002, 1.027) *0.020*	X-158.634	1.013 (1.003, 1.028) *0.018*
Platelet count		X-227.532	0.996 (0.992, 0.999) *0.015*	X-227.532	0.996 (0.992, 0.999) *0.017*
Serum creatinine		X-61.615	1.015 (1.003, 1.026) *0.012*	X-61.615	1.015 (1.003, 1.026) *0.012*
Antihypertensive treatment			1.291 (0.675, 2.469) *0.440*		1.293 (0.677, 2.469) *0.436*
Magnesium sulphate treatment			5.567 (3.005, 10.314) *<0.001*		5.538 (2.994, 10.243) *<0.001*

HR = Hazard Ratio, 95% CI = 95% confidence interval.
[¥]Medical history is a count of the following conditions: chronic hypertension, renal disease, diabetes mellitus, and previous pre-eclampsia.

The Cox and Royston-Parmar regression methods produce almost identical results (Table 4.5). Both models exclude ALT due to a statistically insignificant association with the risk of antenatal adverse events, and both models include all other continuous candidate prognostic factors as linear terms. Maternal age, gestational age, antihypertensive treatment, and magnesium sulphate treatment were forcibly retained in both models for clinical acceptability and to account for treatment effects; however only magnesium sulphate treatment returned a statistically significant association $p < 0.001$. Participants receiving magnesium sulphate treatment appear to have an increased risk of antenatal adverse events (HR 5.567 Cox, 5.538 Royston-Parmar). This treatment is given to minimise the risk of eclampsia and prevent pre-term labour, thus the association is likely to represent the identification and treatment of high risk participants by clinicians. Both models also identified a statistically significant association between the risk of antenatal adverse events and previous medical history ($p = 0.038$ Cox, $p = 0.037$ Royston-Parmar). Surprisingly the risk of antenatal adverse events is reduced for patients with pre-existing medical conditions compared to those with noneThis finding is in line with existing prognostic models currently used for women with pre-eclampsia when predicting maternal adverse events (von Dadelszen et al., 2009, Thangaratinam et al., 2017), thus should not affect the face validity of the model.

4.3.4 Baseline cumulative risk and fitted prognostic models

Baseline cumulative risk estimates from the multivariable prognostic models were obtained through transformation of the estimated baseline cumulative hazard function. Estimates over time are presented graphically in Figure 4.4, and are provided for a number of pre-specified time points in Table 4.6.

Figure 4.4: Estimates of baseline cumulative risk from two multivariable prognostic models

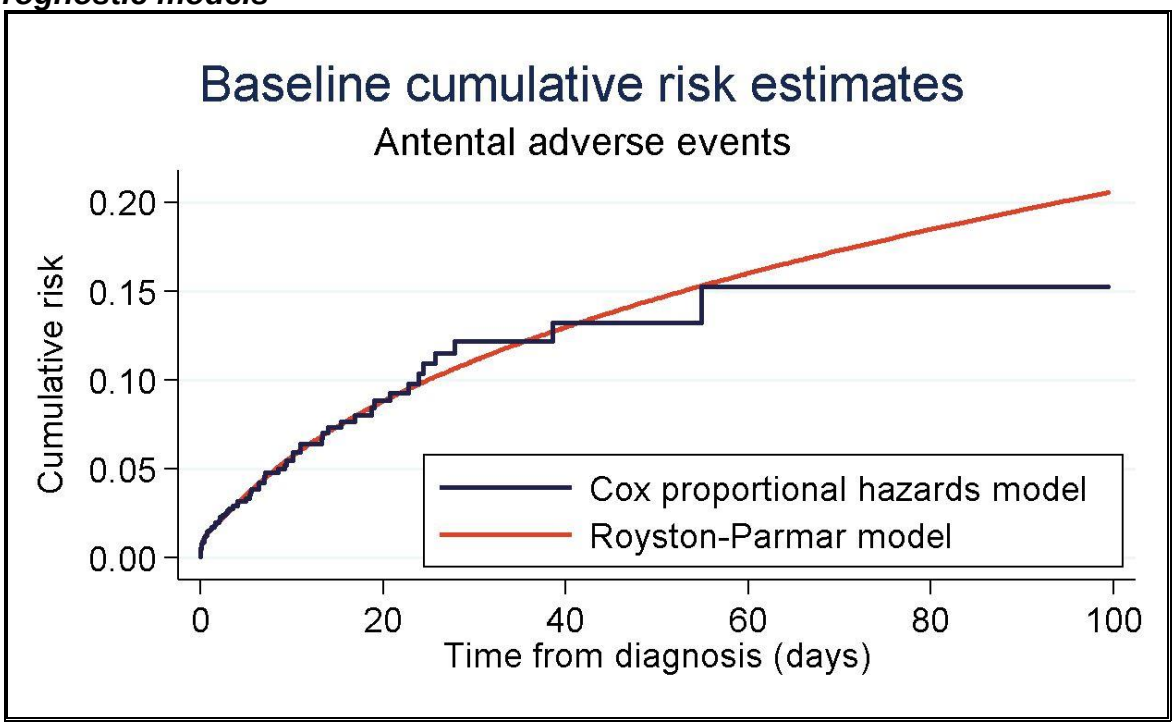


Table 4.6: Estimates of baseline cumulative risk from two multivariable prognostic models at given time points

	2 days	1 week	4 weeks
Cox proportional hazards model	0.015	0.038	0.116
Royston-Parmar model	0.016	0.039	0.102
Difference	0.001	0.001	-0.014

The baseline cumulative risk estimates are calculated using centred values of the transformed prognostic factors; thus, this risk relates to a hypothetical participant with mean continuous prognostic factor values and 0 categorical prognostic factor values. Both the Cox and Royston-Parmar models produce similar baseline survival estimates up to three weeks following diagnosis. Differences between the baseline cumulative incidence estimates from the models become apparent after three weeks, due to the differences in the approaches used by each model to estimate the baseline survival functions. The Cox model requires an additional non-parametric analysis to be performed to calculate the baseline survival function. This approach results in a stepped survival function in which baseline estimates only change at the time of an observed event. As few events occurred after three weeks, the baseline survival

estimates after this time remain fairly static, resulting in a plateau in estimated baseline risk between weeks four and five. The Royston-Parmar model directly models the baseline log cumulative hazard using a smooth spline function.

The fitted prognostic models combine the baseline survival estimates, $S_0(t)$, depicted in Figure 4.4 with the multivariable regression coefficients, $\hat{\boldsymbol{\beta}}$, from Table 4.5 using the equation $S(t) = S_0(t)^{\exp\{\hat{\boldsymbol{\beta}}\mathbf{X}\}}$. The equations for these fitted models (i.e. before adjustment for overfitting) are provided in Table 4.7.

Table 4.7: Equations for fitted prognostic models for antenatal adverse events

Cox proportional hazards model	**Royston-Parmar flexible parametric model**
$Linear\ Predictor\ (\widehat{\boldsymbol{\beta}}\boldsymbol{X}) =$ $- 0.025 \times (\boldsymbol{maternal\ age} - 30.244)$ $+ 0.018 \times (\boldsymbol{gestational\ age} - 30.562)$ $- \begin{cases} 0.783, & if\ \boldsymbol{1\ preexisting\ condition} \\ 0.419, & if\ \boldsymbol{2\ or\ more\ conditions} \end{cases}$ $+ 0.015 \times (\boldsymbol{SBP} - 158.634)$ $- 0.004 \times (\boldsymbol{platelet\ count} - 227.532)$ $+ 0.015 \times (\boldsymbol{serum\ creatinine} - 61.615)$ $+ 0.255 \times (if\ \boldsymbol{antihypertensive})$ $+ 1.717 \times (if\ \boldsymbol{MgSO4\ treatment})$ $S_0(t) = \begin{cases} 0.985, & if\ t = 2\ days \\ 0.962, & if\ t = 1\ \text{week} \\ 0.884, & if\ t = 4\ \text{weeks} \end{cases}$	$Linear\ Predictor\ (\widehat{\boldsymbol{\beta}}\boldsymbol{X}) =$ $- 0.039 \times (\boldsymbol{maternal\ age} - 30.244)$ $+ 0.015 \times (\boldsymbol{gestational\ age} - 30.562)$ $- \begin{cases} 0.777, & if\ \boldsymbol{1\ preexisting\ condition} \\ 0.456, & if\ \boldsymbol{2\ or\ more\ conditions} \end{cases}$ $+ 0.013 \times (\boldsymbol{SBP} - 158.634)$ $- 0.004 \times (\boldsymbol{platelet\ count} - 227.532)$ $+ 0.015 \times (\boldsymbol{serum\ creatinine} - 61.615)$ $+ 0.257 \times (if\ \boldsymbol{antihypertensive})$ $+ 1.712 \times (if\ \boldsymbol{MgSO4\ treatment})$ $S_0(t) = \begin{cases} 0.984, & if\ t = 2\ days \\ 0.961, & if\ t = 1\ \text{week} \\ 0.898, & if\ t = 4\ \text{weeks} \end{cases}$
$Pr(Antenatal\ adverse\ event) = 1 - S_0(t)^{exp(\widehat{\boldsymbol{\beta}}\boldsymbol{X})}$	

4.3.5 Sensitivity analysis of fitted models

A sensitivity analysis was performed to investigate the proportional hazards assumption made by both the Cox and Royston-Parmar modelling approaches. Interactions between the prognostic factors and the natural log of time were incorporated into the fitted models to test this assumption. The results are given in Appendix IX; Significant interactions were detected between magnesium sulphate treatment and ln(time) in both models ($p = 0.009$ Cox, $p < 0.001$ Royston-Parmar), and between SBP and ln(time) in the Royston-Parmar model ($p = 0.010$). The inclusion of

these non-proportional estimates into the final prognostic model could result in further overfitting of the model, due to the small sample size of this study. Additionally, the inclusion of significant time interactions is unlikely to affect the prognostic performance of the models, as the time points assessed are soon after diagnosis (2 days, 1 week, and 4 weeks). Thus, the time interactions were intentionally disregarded from the final models. Not incorporating these interactions is unlikely to affect the results of this chapter, which focuses on the comparison of the two time-to-event modelling approaches.

4.3.6 Apparent performance of fitted prognostic models

The apparent performance of the models was assessed in the imputed PREP study data. The overall calibration of both prognostic models was assessed at two days, one week, and four weeks, the results are displayed in Table 4.8. Both models struggle to predict the brisk increase in risk directly after pre-eclampsia diagnosis and thus overall calibration at early time points (two days and one week) is suboptimal. However, both models appear to calibrate well at later time points. Averaging across multiple datasets, the calibration slope for the Cox proportional hazards model was found to be 0.994 (95% CI: 0.79 to 1.20), whereas the calibration slope for the flexible parametric model was 1.004 (95% CI: 0.80 to 1.21). The multiple imputation methods used to develop these models cause slightly imperfect (not equal to 1.0) calibration slopes, but they are almost perfect, as expected given the model was developed using the same data

Table 4.8: Overall calibration of prognostic models at time points

	Kaplan-Meier observed	Cox proportional hazards model		Royston-Parmar flexible parametric model	
		Expected	E/O	Expected	E/O
2 days	3.6%	1.6%	0.444	1.6%	0.444
1 week	6.2%	3.8%	0.613	3.9%	0.629
4 weeks	12.6%	11.6%	0.921	12.3%	0.976

The discrimination of the models was assessed using Harrell's C-index at two days, one week, four weeks, and overall as well as using the R^2_D measure, obtained from the D-statistic; the results are given in Table 4.9. Both models perform similarly and show good ability to discriminate between those participants who experienced an antenatal adverse event and those who did not.

Table 4.9: Measures of discrimination

	Harrell's C-index				Royston and Sauerbrei's D-statistic	
	2 days	1 week	4 weeks	Overall (95% CI)	D-statistic	R^2_D
Cox model	0.847	0.801	0.784	0.784 (0.723, 0.845)	1.924	0.469
Royston-Parmar model	0.849	0.804	0.786	0.786 (0.726, 0.847)	1.915	0.444

In summary, the apparent prognostic performance of both the models is similar. As expected, both are well calibrated and have good discriminative ability when assessed using the dataset in which they were developed.

4.3.7 Internal calibration and optimism adjustment

The optimism of the prognostic performance of the two prognostic models was assessed through internal validation using 100 bootstrap samples. The internal validation of the models was assessed using measures of calibration (calibration slope)

and discrimination (Harrell's C and R^2_D). The average performance measures from the 100 bootstrap samples are reported in Table 4.10.

Table 4.10: Measures of prognostic performance from 100 bootstrap samples

		Average bootstrap performance	Average test performance	Average optimism	Optimism-adjusted performance
Cox model	**Calibration slope**	1.000	0.837	0.163	0.837
	Harrell's C-index	0.799	0.767	0.031	0.768
	D-statistic	2.044	1.691	0.329	1.715
	R2D Statistic	0.499	0.406	0.089	0.410
Royston-Parmar model	**Calibration slope**	1.000	0.837	0.163	0.837
	Harrell's C-index	0.801	0.767	0.031	0.770
	D-statistic	2.049	1.702	0.336	1.713
	R2D Statistic	0.501	0.409	0.090	0.411

Again, both of the prognostic models perform similarly in terms of prognostic performance and levels of optimism. The average optimism in Harrell's C statistic was found to be 0.031 for both the Cox and Royston-Parmar models, resulting in optimism adjusted C-index of 0.799-0.031=0.768 for the Cox model and 0.801-0.031=0.770 for the Royston-Parmar model. The average optimism in the R^2_D measure was found to be 0.089 for the Cox model and 0.090 for the Royston-Parmar model, resulting in similar optimism adjusted R^2_D measures of 0.499-0.089=0.410 for the Cox model and 0.501-0.090=0.411 for the Royston-Parmar model. Even after adjusting for optimism, the prognostic measures for discrimination for both models are promising. The average optimism in the calibration slope was 0.163 for both the Cox and Royston-Parmar models, resulting in uniform shrinkage factors of 1.000-0.163=0.837 for both models.

4.3.8 Development of final (optimism adjusted) model equations

The fitted models were adjusted for optimism by multiplying the uniform shrinkage factor for each model to all predictor coefficients within the fitted prognostic models.

Then. the baseline cumulated hazard functions were re-estimated with the new linear predictors to ensure calibration-in-the-large. The regression equations for the final optimism adjusted prognostic models are provided in Table 4.11.

Table 4.11: Final regression equations for optimism adjusted prognostic models

Cox proportional hazards model	Royston-Parmar flexible parametric model
$Linaer\ Predictor\ (\widehat{\boldsymbol{\beta}}\boldsymbol{X}) =$ $- \ 0.021 \times (\boldsymbol{maternal\ age} - 30.244)$ $+ \ 0.015 \times (\boldsymbol{gestational\ age} - 30.562)$ $- \begin{cases} 0.661, & if\ \boldsymbol{1\ preexisting\ condition} \\ 0.347, & if\ \boldsymbol{2\ or\ more\ conditions} \end{cases}$ $+ \ 0.012 \times (\boldsymbol{SBP} - 158.634)$ $- \ 0.003 \times (\boldsymbol{platelet\ count} - 227.532)$ $+ \ 0.012 \times (\boldsymbol{serum\ creatinine} - 61.615)$ $+ \ 0.214 \times (if\ \boldsymbol{antihypertensive})$ $+ \ 1.437 \times (if\ \boldsymbol{MgSO4\ treatment})$ $S_0(t) = \begin{cases} 0.980, & if\ t = 2\ days \\ 0.956, & if\ t = 1\ \text{week} \\ 0.878, & if\ t = 4\ \text{weeks} \end{cases}$	$Linaer\ Predictor\ (\widehat{\boldsymbol{\beta}}\boldsymbol{X}) =$ $- \ 0.032 \times (\boldsymbol{maternal\ age} - 30.244)$ $+ \ 0.012 \times (\boldsymbol{gestational\ age} - 30.562)$ $- \begin{cases} 0.650, & if\ \boldsymbol{1\ preexisting\ condition} \\ 0.381, & if\ \boldsymbol{2\ or\ more\ conditions} \end{cases}$ $+ \ 0.011 \times (\boldsymbol{SBP} - 158.634)$ $- \ 0.003 \times (\boldsymbol{platelet\ count} - 227.532)$ $+ \ 0.012 \times (\boldsymbol{serum\ creatinine} - 61.615)$ $+ \ 0.215 \times (if\ \boldsymbol{antihypertensive})$ $+ \ 1.433 \times (if\ \boldsymbol{MgSO4\ treatment})$ $S_0(t) = \begin{cases} 0.979, & if\ t = 2\ days \\ 0.955, & if\ t = 1\ \text{week} \\ 0.893, & if\ t = 4\ \text{weeks} \end{cases}$
$Pr(Antenatal\ adverse\ event) = \ 1 - S_0(t)^{exp(\widehat{\boldsymbol{\beta}}\boldsymbol{X})}$	

4.3.9 Application of final optimism adjusted models to new individuals

The above findings show that the two prognostic modelling approaches gave similar results at each stage of model development and internal validation. In order to compare the predicted risk estimates produced by the final models, a fictitious patient (*Patient Z*) will be used for illustration. *Patient Z* is 25 years old and was diagnosed with pre-eclampsia 33.8 weeks into her pregnancy, she has no pre-existing medical conditions and has received treatment with magnesium sulphate. Their baseline test results are as follows: SBP= 159, platelet count= 226, ALT= 497, serum creatinine= 61. Patient Z's baseline characteristics are displayed in Table 4.12.

In order to calculate *Patient Z*'s risk of experiencing an antenatal adverse event, the values above were incorporated into the final optimism adjusted prognostic model regression equations. The linear predictor equations from the optimism adjusted

models for *Patient Z* are given in Table 4.13. The difference between the linear predictor values from the two models for the hypothetical patient is 0.045.

Table 4.12: Patient Z's baseline characteristics

Prognostic factors	Patient Z
Maternal age (MA)	25
Gestational age (GA)	33.8
Medical history (MH)	0
Systolic blood pressure (SBP)	159
Platelet count (PC)	226
Serum Creatinine (SC)	61
Antihypertensive treatment (AH)	0
Magnesium sulphate treatment (MG)	1

Table 4.13: Linear predictor calculation for Patient Z

Cox proportional hazards model	Royston-Parmar flexible parametric model
$LP =$ $-\ 0.021 \times \left((\boldsymbol{MA} = \mathbf{25}) - 30.244\right)$ $+\ 0.015 \times\ \left((\boldsymbol{GA} = \mathbf{33.8}) - 30.562\right)$ $+\ 0.012 \times \left((\boldsymbol{SBP} = \mathbf{159}) - 158.634\right)$ $-\ 0.003\ \times \left((\boldsymbol{PC} = \mathbf{226}) - 227.532\right)$ $+\ 0.012\ \times \left((\boldsymbol{SC} = \mathbf{61}) - 61.615\right)$ $+\ 1.437\ \times (\boldsymbol{MG} = \mathbf{1})$ $= 1.599$	$LP =$ $-\ 0.032 \times \left((\boldsymbol{MA} = \mathbf{25}) - 30.244\right)$ $+\ 0.012 \times\ \left((\boldsymbol{GA} = \mathbf{33.8}) - 30.562\right)$ $+\ 0.011 \times \left((\boldsymbol{SBP} = \mathbf{159}) - 158.634\right)$ $-\ 0.003\ \times \left((\boldsymbol{PC} = \mathbf{226}) - 227.532\right)$ $+\ 0.012\ \times \left((\boldsymbol{SC} = \mathbf{61}) - 61.615\right)$ $+\ 1.433\ \times (\boldsymbol{MG} = \mathbf{1})$ $= 1.644$

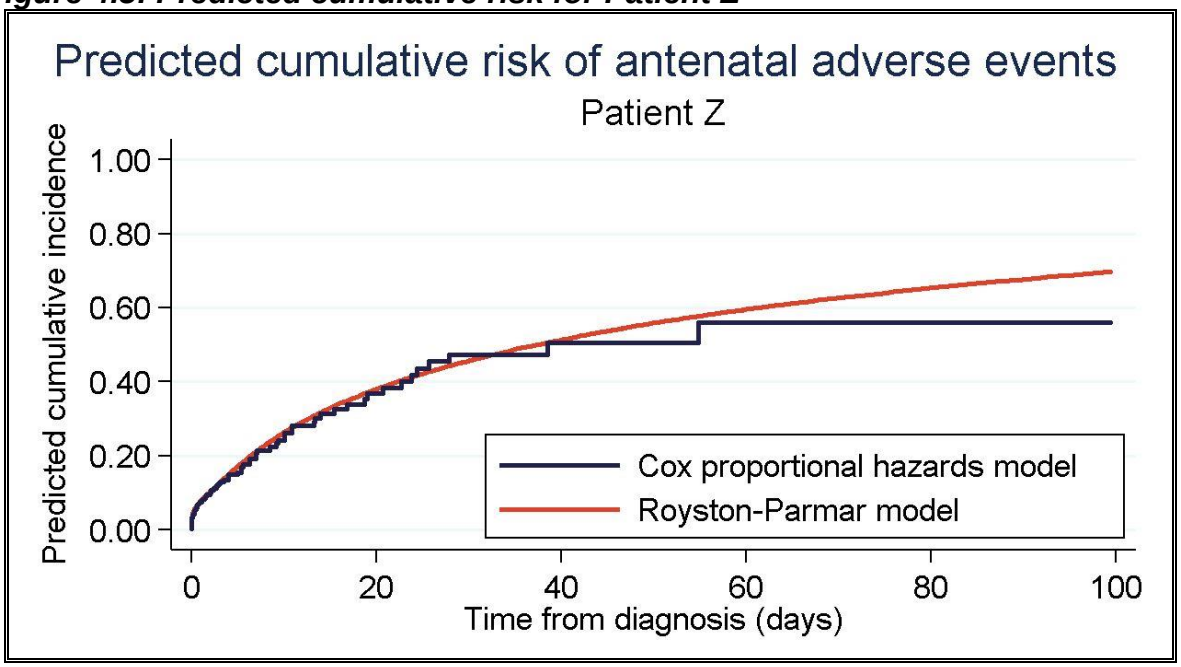
The values from the linear predictor calculations were then combined with the re-estimated baseline survival estimates to give the predicted cumulative risk of experiencing an antenatal adverse event at two days, one week, and four weeks for *Patient Z*. The calculations for predicted risks are shown in Table 4.14.

The difference between the predicted risk estimates from the two models becomes larger as time progresses. A graphical display of *Patient Z*'s predicted cumulative risk of an antenatal event over time is depicted in Figure 4.5.

Table 4.14: Predicted cumulative risk of antenatal adverse events for Patient Z

Cox proportional hazards model	Royston-Parmar flexible parametric model
$Risk = 1 - S_0(t)^{exp(1.599)}$	$Risk = 1 - S_0(t)^{exp(1.644)}$
$1 - 0.980^{\exp(1.599)} = 9.5\%, by\ 2\ days$	$1 - 0.979^{\exp(1.644)} = 10.4\%, by\ 2\ days$
$1 - 0.956^{\exp(1.599)} = 20.0\%, by\ 1\ week$	$1 - 0.955^{\exp(1.644)} = 21.2\%, by\ 1\ week$
$1 - 0.878^{\exp(1.599)} = 47.5\%, by\ 4\ weeks$	$1 - 0.893^{\exp(1.644)} = 44.3\%, by\ 4\ weeks$
$1 - 0.868^{\exp(1.599)} = 50.4\%, by\ 40\ days$	$1 - 0.870^{\exp(1.644)} = 51.4\%, by\ 40\ days$
$1 - 0.847^{\exp(1.599)} = 56.0\%, by\ 60\ days$	$1 - 0.840^{\exp(1.644)} = 59.4\%, by\ 60\ days$
$1 - 0.847^{\exp(1.599)} = 56.0\%, by\ 80\ days$	$1 - 0.816^{\exp(1.644)} = 65.1\%, by\ 80\ days$

Figure 4.5: Predicted cumulative risk for Patient Z



Both models appear to give quite similar predictions of cumulative risk up to around 40 days, however after this time the model predictions diverge slightly. For example, at 60 days the predicted risk is 56.0% from the Cox model but 59.4% from the Royston-Parmar model. Arguably, the estimates over time produced by the Royston-Parmar model provide a more realistic risk profile for a patient compared to those produced by the Cox model. It is much more likely that the real world risk of an event is smoothly increasing over time, rather than jumping at specific time points (when an event is observed in another participant) and remaining stable in between these time points.

4.4 Discussion

In this chapter, both Cox and Royston-Parmar regression methods were applied to develop two new prognostic models which predict the risk of antenatal adverse events in women diagnosed with early-onset pre-eclampsia. The models were internally validated and adjusted for optimism; the included prognostic factors, predictive performance, and final optimism adjusted equations were compared. The aim of this chapter was to demonstrate the standard time-to-event methods to develop prognostic models and compare models developed using two popular time-to-event approaches.

Two commonly used time-to-event modelling methods, the Cox proportional hazards method (Cox, 1972a) and Royston-Parmar flexible parametric method (Royston and Parmar, 2002), were applied to develop prognostic models. Though the two modelling methods produced practically identical regression coefficient estimates, the difference between the estimation methods for the baseline survival function resulted in differences in predicted risks for individual patients over time, especially at later time points. The Royston-Parmar flexible parametric modelling approach directly estimates a smooth and flexible baseline cumulative hazard function, which can be utilised to produce smooth predicted risk estimates for study participants. This is advantageous over the Cox model, which requires additional modelling and estimation to estimate a non-parametric step function for the baseline cumulative hazard function. This stepped function becomes problematic at later time points as the number of participants and events decreases and the steps grow in size. At later time points the baseline hazard function from the Royston-Parmar approach extrapolates with little data, but the smooth function is more realistic. Hence, it is recommended the flexible parametric approach be used in prognostic model research and will be the focus of the remainder of this thesis.

4.4.1 Limitations and further research

This chapter adds to the work conducted by the PREP study team (Thangaratinam et al., 2017) in developing prognostic models which may be used to inform pre-eclampsia patients of their risks of adverse events. The alteration of the definition of the outcome to *antenatal* adverse events compliments the prognostic models developed by the PREP study team, and both models may be used to give the patient a broader picture of their risk of events during pregnancy and their risk of preterm delivery. However, changing the outcome does not completely eradicate the risk of treatment paradox bias (Cheong-See et al., 2016), as patients were censored at the time of delivery of the baby. Therefore, the risk predictions calculated using the prognostic models developed in this chapter should be interpreted as the risk of experiencing an antenatal adverse event over time in a hypothetical scenario where delivery of the baby (even for treatment) is not possible.

Other methods have been proposed to reduce the risk of treatment paradox bias including: deleting the “trigger” prognostic factor from the model (Schuit et al., 2013), standardising treatment across predictor levels, using propensity scores (Cheong-See et al., 2016), and modelling the probability of treatment (Groenwold et al., 2016). However, considering the definition of a competing event (an event which prevents or alters the risk of the event of interest from occurring (Koller et al., 2012)), a reasonable approach might be to model the treatment as a competing risk to the event of interest. By incorporating competing risks methodology into the development of prognostic models, one would be able to produce “real-world” risk prediction which would account for the presence and probability of the treatment happening. In the following chapter, the statistical methods to appropriately account for competing events will be utilised to develop prediction models for the risk of antenatal adverse events, accounting for the competing risk of delivery. Later, in Chapter 6, an additional dataset will be utilised to externally validate and compare the standard time-to-event and competing risks models developed in this and the following chapter.

5 A COMPARISON OF PROGNOSTIC MODELS DEVELOPED USING FLEXIBLE PARAMETRIC COMPETING RISKS METHODS

5.1 Introduction

In this chapter, two prognostic models developed using different flexible parametric competing risks approaches (introduced in Section 1.4.5), namely the cause-specific (Hinchliffe and Lambert, 2013b) and subdistribution (Lambert et al., 2017) approaches are applied to develop two new prognostic models. Data from the PREP study (Thangaratinam et al., 2017), introduced in Chapter 4, will be used to develop and internally validate the models to predict the risk of antenatal adverse events in women diagnosed with early-onset pre-eclampsia, accounting for the competing event of delivery. The resulting models will be compared to illustrate similarities and differences between the two competing risks approaches.

5.1.1 Prediction of antenatal adverse events with competing events

The original PREP study (Thangaratinam et al., 2017) developed two multivariable prognostic models to predict the risk of adverse maternal outcomes in women diagnosed with confirmed early-onset pre-eclampsia. Pre-term delivery was included in the composite outcome, which resulted in limitations in the interpretation of the model predictions. Some of these limitations were addressed through the development of prognostic models which predict the risk of an alternative outcome, namely antenatal adverse events (Chapter 4). The models developed in Chapter 4 censor participants following delivery of the baby, as antenatal adverse events occur during pregnancy and delivery of the baby ends

the pregnancy. However, as delivery of the baby prevents antenatal adverse events from occurring, it is consequently a competing event and should be managed appropriately to avoid competing risks bias. Thus competing risks methods should be utilised in the development of prognostic models for antenatal adverse events.

5.1.2 Aims

This chapter expands on the analysis performed in Chapter 4 to appropriately account for the competing event of delivery when developing prognostic models for antenatal adverse events in women with early-onset pre-eclampsia. The developed models are presented, internally validated, and adjusted for optimism. The parameter estimates of the prediction models will be compared, as will the resulting model predictions, measures of prognostic ability, and optimism.

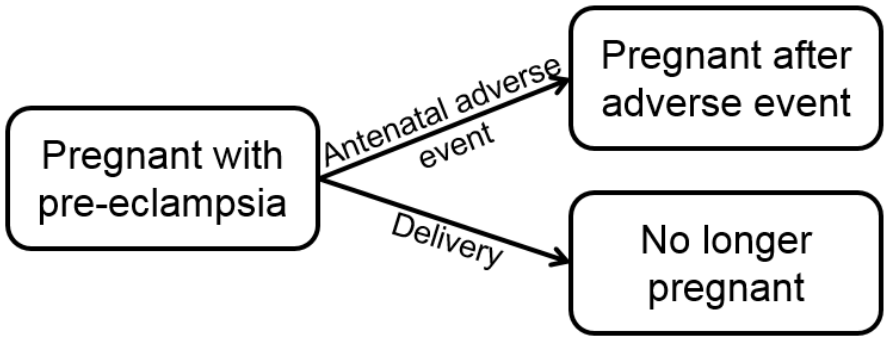
5.2 Methods: Developing two prognostic models with competing events

Methods for the development and internal validation of the flexible parametric competing risks prognostic models are now described.

5.2.1 Antenatal adverse events with competing risks outcome definition

The time to an antenatal adverse event in women with pre-eclampsia can be considered as a multi-state structure, as depicted in Figure 5.1. Participants enter the initial state on the date of a confirmed diagnosis of pre-eclampsia, and remain in this state until they experience an event. On the date an event occurs, the patient transitions out of the initial state into one of the two final absorbing states.

Figure 5.1: Competing risks multi-state structure for PREP study



Boxes represent states and arrows represent transitions.
All participants enter the study at pre-eclampsia diagnosis (the initial state).
Following an antenatal adverse event event (the event of interest, k=1), participants transition into the associated absorbing state (pregnant following adverse event).
Participants who deliver the baby (the competing event) transition into a different absorbing state (no longer pregnant).

The definition of an antenatal adverse event remains unchanged from that outlined in Section 4.2.1; a composite outcome comprised of component events listed in Table 5.1. However, in this chapter delivery of the baby, which prevents the event of interest from occurring, is now considered a competing event (recall deliveries were censored in the previous chapter).

Table 5.1: Individual components of antenatal adverse events

Antenatal adverse events
Placental abruption, transfusion of blood, eclamptic seizure, hepatic dysfunction, pulmonary oedema, intubation, acute renal insufficiency, at least 50% FIO2 for > 1hour, Glasgow Coma score <13

The primary aim of the prognostic models developed in this chapter is to predict the risk of experiencing an antenatal adverse event, which by definition occurs prior to delivery. The use of competing risks methods will allow the calculation of this risk while appropriately accounting for the presence of the competing event (delivery of the baby). Though delivery of the baby can occur *after* an antenatal adverse event, it is not necessary to model this transition when evaluating the risk of antenatal adverse events which occur *before* delivery. Events which occur *after* delivery, such as a postpartum haemorrhage, are again not captured by this structure so are excluded from the definition. If both an adverse event and delivery occurred on the same date the outcome was recorded as an antenatal adverse event, as it is likely that this adverse event would trigger induction of labour as a treatment.

5.2.2 Candidate prognostic factors

The restricted set of candidate prognostic factors considered in Chapter 4 (listed in Table 5.2) was utilised during the development of the competing risks prognostic models.

Table 5.2: Candidate prognostic factors for competing risks prognostic models

Maternal characteristics: Maternal age at diagnosis (years), Gestational age at diagnosis (weeks).
Medical history: Count of pre-existing conditions (0, 1, 2 or more) from pre-specified list: pre-existing hypertension, renal disease, diabetes mellitus, previous history of pre-eclampsia.
Bedside examination and tests: Systolic blood pressure (mmHg, highest measurement over 6 hrs), Platelet count (x 109/L), Alanine amino transaminase (IU/l), Serum creatinine µmol/L.
Management at baseline (before or within 1 day of diagnosis): Administration of oral and/or parenteral anti-hypertensives, Administration of magnesium sulphate.

5.2.3 Competing events risk sets and data structures

The two competing risks approaches require the study data to be re-structured to reflect the associated risk set (depicted in Figure 1.9). The cause-specific approach required the data to be restructured using the multi-state structure (see Figure 1.11), with multiple rows for each participant. The subdistribution approach required those participants who experienced a competing event to remain in the risk set (see Figure 1.9). Thus, the time-to-event for those who delivered was replaced with the time of the last observed antenatal adverse event in the entire dataset (t=117 days), at which time these participants were censored.

5.2.4 Non-parametric assessment of competing risks bias

Competing risks bias was assessed by comparing the non-parametric Kaplan-Meier curve which censor the competing event of delivery (shown in Figure 4.1) to non-parametric cumulative incidence curves which appropriately account for competing events, estimated using the subdistribution risk set. The difference between these curves represents the amount of competing risks bias present when the competing events are not appropriately accounted for.

5.2.5 Multiple imputation with competing events

Though complete data were available for a large proportion (829, 87.5%) of the PREP cohort, multiple imputation using chained equations (Van Buuren et al., 1999) techniques were applied to missing information to avoid a loss of information. Under a missing at random assumption, a total of 20 imputed datasets were created from the original (not restructured) PREP study data using the *mi impute chained* package in Stata 14. The chained equations included the candidate prognostic factors listed in Table 5.2, with medical history separated into its individual components. The appropriate outcome information for each competing risk approach (discussed in Section 1.5.1) was also incorporated into the chained equations. Missing values from continuous prognostic factors were imputed using the predictive mean matching method (Little, 1988), and those from binary prognostic factors were imputed using augmented logistic regression (White et al., 2010). The imputed datasets were then restructured to reflect the competing event risk sets as described in 5.2.3. Imputed values were visually inspected to determine any outliers, but none were identified. Information on outcome measures, including pre-eclampsia diagnosis date, date of event, and event type, were complete and thus were not imputed.

5.2.6 Flexible parametric competing risks regression models

Two new flexible parametric prognostic models were developed to predict the risk of antenatal adverse events in women diagnosed with early onset pre-eclampsia accounting for the competing event of delivery. The first was developed using the cause-specific approach; the second was developed using the subdistribution approach. Details of the statistical methods of these approaches, provided in more detail in Section 1.4.5, are summarised below.

Briefly, the cause-specific approach (Hinchliffe and Lambert, 2013b) was used to develop one flexible parametric competing risks model using the *multistate* package with the *stpm2* estimation command in Stata. Restricted cubic splines of log-time are

utilised to approximate the log cumulative baseline cause-specific hazard function. The model regression coefficients and spline function parameter estimates are calculated simultaneously through maximum partial likelihood techniques. Estimates of both the cause-specific hazard function for antenatal adverse events and the cause-specific hazard function for delivery are required to obtain an estimate of the cause-specific cumulative incidence function (i.e. risk of having an antenatal adverse event over time). Plots of estimated cumulative cause-specific hazard functions for null models (containing no prognostic factors), with splines of varying degrees of freedom (between two and five), were inspected for each event separately to determine the number of knots required to capture the shape of the underlying functions.

The subdistribution approach (Lambert et al., 2017) was used to develop the other flexible parametric competing risks model using the *stpm2* estimation command in Stata. Restricted cubic splines of log-time are utilised to approximate the log cumulative baseline subdistribution hazard function. The model regression coefficients and spline function parameter estimates are calculated simultaneously through maximising the weighted partial likelihood function (Lambert et al., 2017). This approach only requires the estimation of the subdistribution hazard function for antenatal adverse events to obtain an estimate of the cause-specific cumulative incidence function (i.e. risk of having an antenatal adverse event over time). The number of knots required to capture the shape of the underlying subdistribution hazard function was determined by visual inspection of the plots of the estimated cumulative subdistribution hazard function for null models (containing no prognostic factors) with splines of varying degrees of freedom (between two and five).

Fractional polynomial terms were incorporated into the models to investigate non-linear associations between continuous prognostic factors and the outcome. The methods for incorporating these terms are described in Section 4.2.6.

5.2.7 Multivariable analysis and prognostic factor selection procedure of prognostic models with competing events

Multivariable analyses, using either the cause-specific or subdistribution approaches, were performed to develop two flexible parametric competing risks prognostic models. A backwards elimination procedure was applied to determine which of the candidate prognostic factors would be included in the final multivariable prognostic models. The procedure utilised the multiply imputed data adapted to incorporate fractional polynomial terms using the *mfpmi* Stata package (Morris et al., 2015). This procedure was additionally adapted to the competing risks approaches as follows:

- The cause-specific approach began with the stacked imputed and restructured cause-specific dataset. The full model was developed using a multi-state modelling approach which simultaneously models the cause-specific hazards for both the event of interest (antenatal adverse events) and the competing event (delivery). The full model incorporated the cause-specific associations of all candidate prognostic factors with each outcome, with all continuous candidate prognostic factors included as linear terms.
- The subdistribution approach began with the stacked imputed and restructured subdistribution dataset. The full model was developed using the subdistribution risk set which models the subdistribution hazards for antenatal adverse events only. The full model incorporated the subdistribution associations for all candidate prognostic factors and the outcome of interest, with all continuous prognostic factors included as linear terms.

The same backwards selection procedure outlined in Box 4.1 was applied to determine which candidate prognostic factors would be retained in the final models. Both maternal and gestational age were retained in the models regardless of their

significance to ensure clinical acceptability of the final models in line with the original PREP study (Thangaratinam et al., 2017). The treatment variables (antihypertensive and magnesium sulphate treatment) were likewise retained to reduce the risk of suboptimal prognostic performance (Groenwold et al., 2016). The multivariable model cause-specific or subdistribution hazard ratios, 95% confidence intervals, and prognostic factor selection results (including exclusion and transformations of the included factors) were compared across the two approaches.

5.2.8 Baseline cumulative incidence functions

The risk of experiencing an antenatal adverse event over time for a participant with “average” prognostic factor values was estimated using the baseline (all prognostic factors centred and set to zero) cause-specific cumulative incidence function. Baseline cause-specific cumulative incidence functions for antenatal adverse events were calculated using both the cause-specific and subdistribution approaches. Both approaches estimate the cumulative baseline cause-specific or subdistribution hazards for the event of interest using the restricted cubic spline function of log time (Equations 1.28 & 1.31). The cause-specific approach also derives an estimate of the cumulative baseline cause-specific hazard for the competing event. These estimates are then transformed to return baseline cause-specific cumulative incidence function estimates. Median baseline cause-specific cumulative incidence estimates for each approach were calculated at two days, one week, and four weeks for patients in the PREP study using the flexible parametric competing risks prognostic models.

5.2.9 Sensitivity analysis

A sensitivity analysis was performed to investigate the assumptions made during model development, namely proportional hazards assumptions and independent effects (i.e. no interactions). Both competing risks approaches when developed in the way described above, assume proportional cause-specific or subdistribution hazards.

That is that the corresponding hazard ratios of the prognostic factors contained within the model remain constant over time, which cannot hold for both models simultaneously. These assumptions were tested for each of the retained prognostic factors in turn by incorporating an interaction between the prognostic factor and time into the final multivariable models. Other clinically plausible interactions between the retained prognostic factors were also investigated in turn to test the assumption that the effects of each prognostic factor on the outcome were independent. Any statistically significant $(\mathrm{P} < 0.05)$ deviations from the assumptions were reported.

5.2.10 Comparison of individual absolute risk predictions

Each participant's individual risk of experiencing an adverse antenatal event was calculated using the cause-specific cumulative incidence function. The cause-specific approach requires cause-specific hazard estimates for both outcomes to return absolute risk estimates (Equation 1.30). Cause-specific cumulative incidence estimates were obtained using the simulation approach (described in Section 1.4.5), which combines estimated regression coefficients with a transition probability matrix (estimated using the cumulative cause-specific hazard function). The subdistribution approach estimates the cause-specific cumulative incidence through the transformation of the combined regression coefficient estimates and the log cumulative subdistribution hazard estimates (Equation 1.32). The predicted risks of antenatal adverse events were calculated for each of the PREP study participants using the flexible parametric competing risks prognostic models. The predicted risks for participants with missing data were calculated using the median imputed value for each prognostic factor for that participant. Median cause-specific cumulative incidence estimates, and the distribution of predicted risk in the PREP study participants, were calculated at two days, one week, and four weeks, these were compared across the two approaches.

5.2.11 Apparent prognostic performance of the fitted prognostic models

Four risk groups were created by splitting the PREP study participants into four equal groups using the 25th, 50th, and 75th centiles of the estimated cause-specific cumulative incidence estimates at four weeks for each approach. The median predicted cause-specific cumulative incidence at two days, one week, and four weeks was calculated for each group, estimates were compared between the modelling approaches. Average predicted cause-specific cumulative incidence functions for each group over time were calculated and compared to non-parametric cumulative incidence estimates.

The apparent performance of the fitted models was assessed in the imputed PREP study data using measures of calibration (overall calibration and calibration slope) and discrimination (Harrell's C-index, D-statistic, and R^2_D), assessed at two days, one week, and four weeks. Overall calibration was examined by comparing the expected number of events (from the prognostic models) to the observed number of events (non-parametric cause-specific cumulative incidence estimates) from the aforementioned risk groups at pre-specified time points. Calibration plots, which graph the expected vs. observed cause-specific cumulative incidence estimates, were utilised to estimate the calibration slope at each of the given time points. The calibration slope was estimated by regressing the individual expected risks (obtained from the prognostic models) on the individual observed outcomes (obtained from non-parametric cumulative incidence estimates), with a Lowess smoother. Harrell's C-index (Harrell et al., 1982) was calculated at the pre-specified time points, utilising the subdistribution risk set to determine which pairs are concordant. Finally, the D-statistic and R^2_D measures were calculated for the models. All measures were calculated in each imputed dataset separately (imputation-specific measures) and were then combined using Rubin's rules. The apparent performance measures for both competing risks prognostic models were compared.

5.2.12 Internal validation and optimism adjustment: subdistribution approach

The model developed using the cause-specific approach was not internally validated, nor adjusted for optimism. The time consuming[24], computationally intensive nature of the simulation approach for obtaining individual risk predictions from this model, alongside the lack of methods for optimism adjustment, resulted in the decision to not proceed further with this particular model. However, the optimism in the apparent performance of the subdistribution model was assessed through internal validation using 100 bootstrap samples of the original (not restructured) PREP study data, as in Chapter 4. The bootstrap validation procedure is outlined in Box 4.2. Optimism adjusted prognostic performance measures were calculated by subtracting the average optimism from the original apparent performance measures as estimated for the fitted model previously (see Section 5.2.11). An optimism adjusted subdistribution model was created by multiplying the fitted regression coefficients by the optimism adjusted calibration slope (representing a uniform shrinkage factor) obtained from the bootstrap procedure. Then, the cumulative baseline subdistribution hazard function was re-estimated to maintain overall calibration between observed and predicted risks.

[24] Fitting the multivariable cause-specific model for death (methods in 5.2.7) took 2 mins 42.8 secs, the multivariable cause-specific model for antenatal adverse events took 3 mins 14.7 secs. Obtaining individual absolute risk predictions for the PREP study participants after the models were fitted took a further 1 hr 17 mins 46.6 seconds.

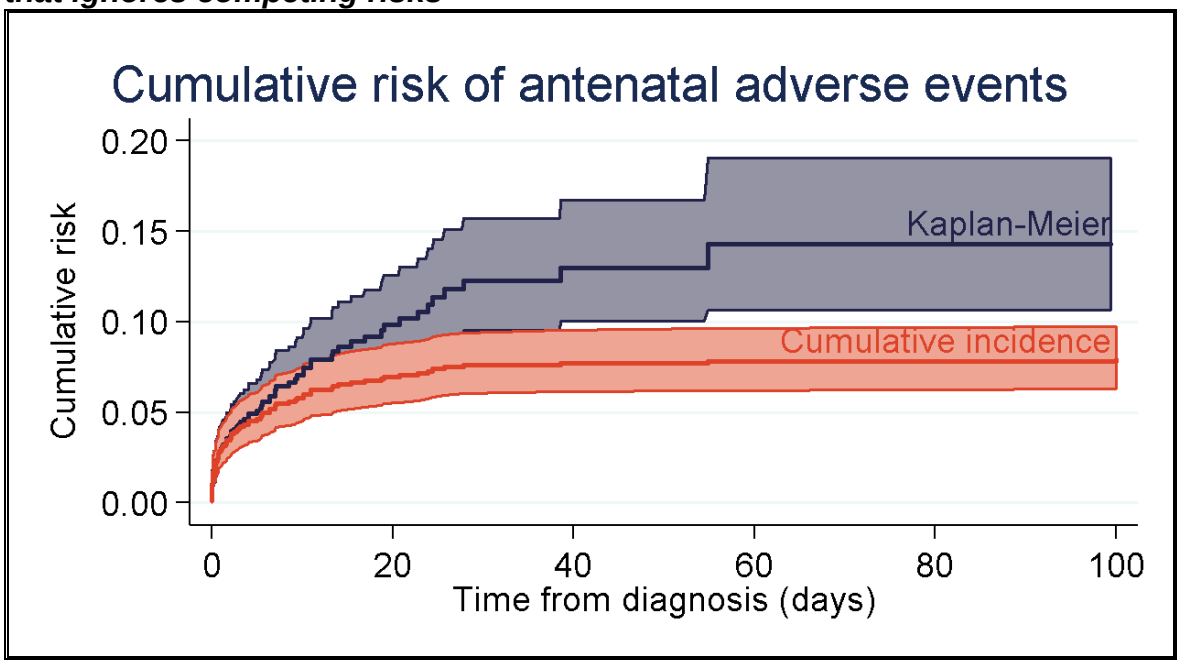
5.3 Results: Comparison of prognostic models for antenatal adverse events developed using flexible parametric competing risks methods.

The PREP study (Thangaratinam et al., 2017) data contained 947 women with a confirmed diagnosis of pre-eclampsia, of whom 75 (7.9%) experienced an antenatal adverse event (i.e. before delivery) and the remaining 872 (92.1%) delivered the baby without an antenatal adverse event.

5.3.1 Non-parametric assessment of competing risks bias

Non-parametric cumulative incidence estimates for antenatal adverse events, which appropriately accounts for competing events, are compared to Kaplan-Meier failure estimates over time: these are depicted in Figure 5.2.

Figure 5.2: Non-parametric cumulative incidence estimates for antenatal adverse events after accounting for competing risks, compared to Kaplan-Meier curve that ignores competing risks

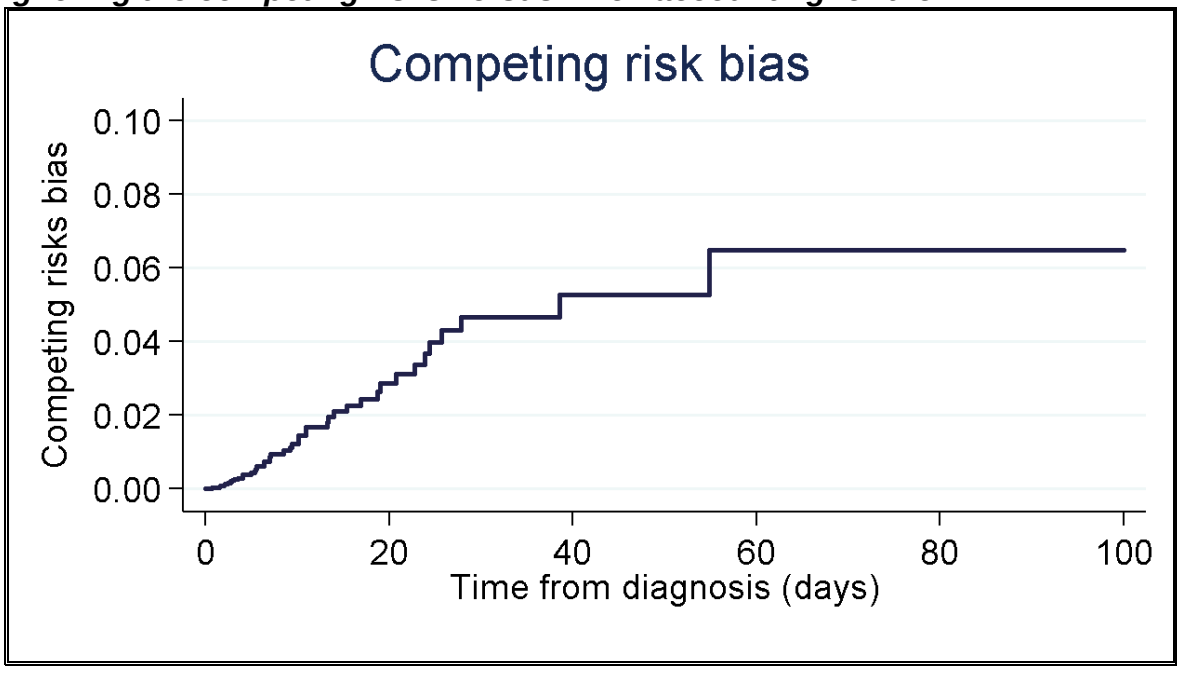


By 100 days, 74 of the 947 women had experienced an antenatal adverse event. As there is no censoring in the study (all participants either experience the event of interest or the competing event) the cumulative incidence at 100 days is $\mathrm{F}_{\mathrm{AAE}}(\mathrm{t}=100)=$

$^{74}/_{947} = 7.8\%$. Whereas the Kaplan-Meier estimate for the risk of experiencing an antenatal adverse event at 100 days is 14.3%. Those who experience competing events are removed from the denominator (risk set) of the Kaplan-Meier estimate but retained for the cumulative incidence estimate. Thus Kaplan-Meier risk estimates are larger than the cumulative incidence estimates and inflate/overestimate the absolute risk of the event of interest in the presence of competing events.

The difference between the Kaplan-Meier and cumulative incidence estimates represents competing risks bias (the difference between ignoring and correctly accounting for competing events). The amount of competing risks bias increases over study time as more competing events occur, this is displayed in Figure 5.3. At 100 days the competing risks bias is 0.065. This may appear small, however when compared to the cumulative incidence estimate (0.079 at 100 days) indicates an 83.1% increase in the absolute risk estimate. Therefore, ignoring the competing events leads to a risk estimate that is nearly double that of the estimated risk when accounting for the competing events.

Figure 5.3: Absolute measure of competing risks bias: difference in non-parametric estimates of cumulative risk of antenatal adverse events when ignoring the competing risks versus when accounting for them



5.3.2 Estimation of baseline cumulative hazard functions

Null models (containing no prognostic factors) were fitted, using spline functions with between two and five degrees of freedom, to determine the number of knots required to adequately capture the shape of the underlying hazard functions. For the subdistribution approach, the cumulative subdistribution hazard function for antenatal adverse events was examined. Whereas for the cause-specific approach, the cumulative cause-specific hazard functions for both antenatal adverse events and delivery were examined separately. The resulting knot locations, Akaikes information criterion (AIC), and Bayesian information criterion (BIC) values from the models are provided in Table 5.3.

Table 5.3: Knot selection for Royston-Parmar flexible parametric model

Approach	Degrees of freedom	Number of knots	Internal knot locations (time in days)	AIC	BIC
Subdistribution appraoch: antenatal adverse events	2	1	2.7	886.9	901.4
	3	2	0.7, 7.0	882.0	901.4
	4	3	0.4, 2.7, 10.1	878.2	902.4
	5	4	0.1, 1.3, 5.4, 13.3	878.7	907.8
Cause-specific appraoch: antenatal adverse events	2	1	2.7	803.4	818.0
	3	2	0.7, 7.0	804.7	824.1
	4	3	0.4, 2.7, 10.1	805.3	829.6
	5	4	0.1, 1.3, 5.4, 13.3	806.1	835.2
Cause-specific approach: delivery	2	1	9.3	3054.4	3068.9
	3	2	4.5, 17.6	3028.8	3048.2
	4	3	3.3, 9.3, 23.4	3029.2	3053.4
	5	4	2.4, 6.3, 14.0, 29.3	3030.3	3059.4

For the subdistribution approach, the spline with three internal knots (4 d.f.) gives the smallest AIC values whereas the splines with one and two internal knots (2 & 3 d.f.) give the smallest BIC values. For the cause-specific approach, the spline with one

internal knot (2 d.f.) gives the lowest AIC and BIC values for antenatal adverse events model, and the spline with two internal knots (3 d.f.) gives the lowest AIC and BIC values for the delivery model.

A graphical display of the spline estimates for the cumulative subdistribution and cause-specific hazards are depicted in Figure 5.4. Using the subdistribution approach, there is little observable difference between the curves with three, four, and five degrees of freedom. Therefore the spline with three degrees of freedom (green line) was considered suitable for the remainder of the analysis. Using the cause-specific approach, there is little observable difference between the spline functions for antenatal adverse events, thus two degrees of freedom was considered suitable for the spline function in the analysis to follow. No notable differences were found between the spline functions for delivery with three, four, and five degrees of freedom, thus three degrees of freedom was considered adequate for prognostic model development. The selected spline functions were compared to non-parametric Nelson-Aalen type cumulative subdistribution and cause-specific hazard estimates (displayed in Figure 5.5), and appear to adequately fit the underlying hazard functions.

Figure 5.4: Comparison of spline functions with varying degrees of freedom for the cumulative subdistribution and cause-specific hazard functions

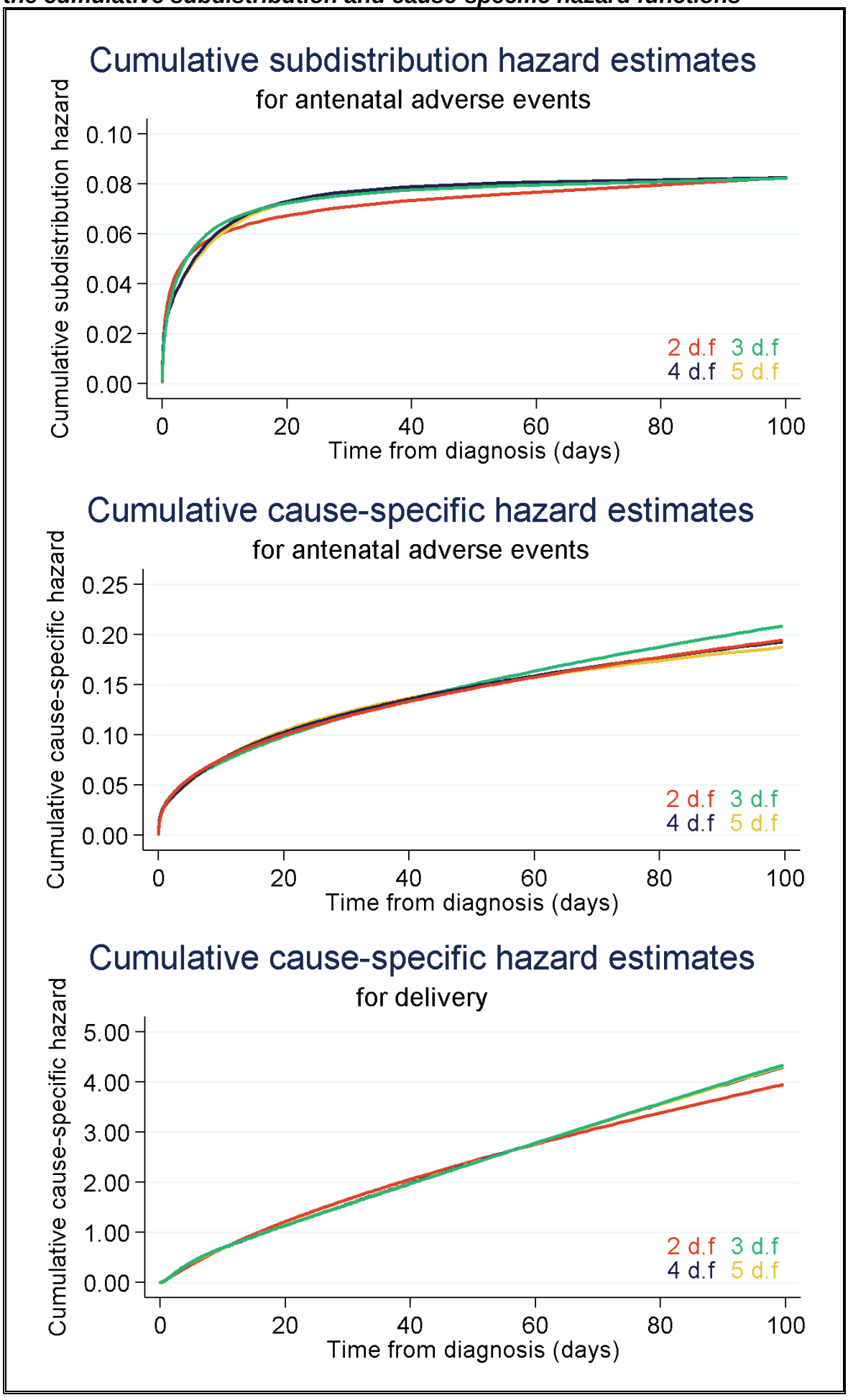
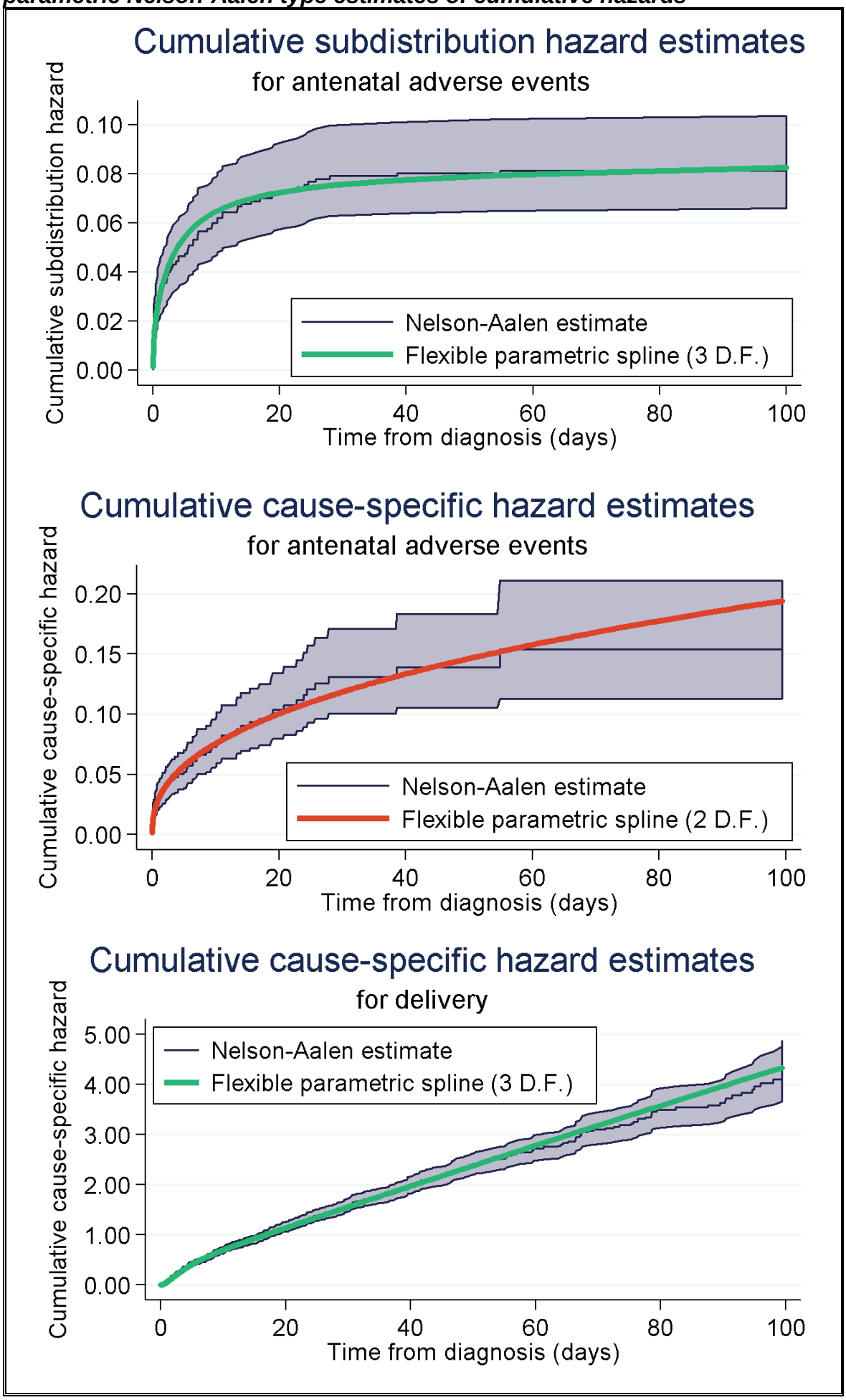


Figure 5.5: Comparison of spline functions selected for modelling against non-parametric Nelson-Aalen type estimates of cumulative hazards



5.3.3 Multivariable analysis and prognostic factor selection procedure of prognostic models for competing events

The backwards elimination procedure for multiply imputed data, incorporating fractional polynomial terms and adapted for competing risks, was applied using both the cause-specific and subdistribution approaches. The resulting cause-specific and subdistribution hazard ratio estimates, 95% confidence intervals and p-values are reported in Table 5.4.

Using the cause-specific approach, all of the candidate prognostic factors were found to be significantly $(p<0.15)$ associated with delivery (the competing event), thus none were excluded from that model. Applying this approach to antenatal adverse events (the event of interest) found medical history, systolic blood pressure, platelet count, serum creatinine, and magnesium sulphate to be significantly $(p<0.15)$ associated with the event. Though the remaining candidate prognostic factors were found not to be statistically significant $(p<0.15)$, maternal age, gestational age, and antihypertensive treatment were forcibly retained in the model. Thus only alanine amino transaminase was excluded from the cause-specific model for antenatal adverse events.

Using the subdistribution approach, only medical history was found to be significantly $(p<0.15)$ associated with antenatal adverse events. However, maternal age, gestational age, and the two treatment prognostic factors were forcibly retained in the model, despite not being statistically significant $(p<0.15)$. The remaining prognostic factors were excluded from this model.

Table 5.4: Multivariable hazard ratios for included prognostic factors

		Cause-specific model: delivery		Cause-specific model: antenatal adverse events		Subdistribution model: antenatal adverse events	
		Transformation of prognostic factor (X)	**$_{CS}$HR (95% CI) *p-value***	**Transformation of prognostic factor (X)**	**$_{CS}$HR (95% CI) *p-value***	**Transformation of prognostic factor (X)**	**$_{SD}$HR (95% CI) *p-value***
Maternal age (years)		X - 30.245	0.973 (0.962, 0.984) *<0.001*	X - 30.245	0.975 (0.938, 1.013) *0.189*	X - 30.245	0.993 (0.955, 1.032) *0.709*
Gestational age (weeks)		((X/10)-2) – 0.107	0.000 (0.000, 0.000) *<0.001*	X - 30.562	1.017 (0.941, 1.098) *0.676*	X - 30.562	0.951 (0.88, 1.028) *0.205*
Medical history*	**1**		0.813 (0.692, 0.955)		0.465 (0.245, 0.882)		0.536 (0.284, 1.011)
	2 or more		0.662 (0.522, 0.84) *0.001*		0.636 (0.275, 1.468) *0.040*		0.830 (0.369, 1.871) *0.128*
Systolic blood pressure		X – 158.64	1.018 (1.014, 1.022) *<0.001*	X – 158.64	1.014 (1.002, 1.027) *0.023*		*Excluded*
Platelet count		X – 227.528	0.997 (0.996, 0.998) *<0.001*	X – 227.528	0.996 (0.993, 0.999) *0.017*		*Excluded*

	Cause-specific model: delivery		Cause-specific model: antenatal adverse events		Subdistribution model: antenatal adverse events	
	Transformation of prognostic factor (X)	**$_{CS}$HR (95% CI) *p-value***	**Transformation of prognostic factor (X)**	**$_{CS}$HR (95% CI) *p-value***	**Transformation of prognostic factor (X)**	**$_{SD}$HR (95% CI) *p-value***
Alanine amino transaminase	Ln(X/1000) +3.446	1.212 (1.095, 1.342) *<0.001*		*Excluded*		*Excluded*
Serum creatinine	X – 61.61	1.013 (1.009, 1.017) *<0.001*	X – 61.61	1.015 (1.003, 1.026) *0.011*		*Excluded*
Antihypertensive treatment		1.343 (1.134, 1.592) *0.001*		1.278 (0.67, 2.439) *0.457*		1.194 (0.633, 2.252) *0.653*
Magnesium sulphate treatment		3.199 (2.553, 4.008) *<0.001*		5.294 (2.891, 9.692) *<0.001*		3.136 (0.633, 2.252) *0.584*

$_{CS}$HR = cause-specific hazard ratio, $_{SD}$HR = subdistribution hazard ratio, 95% CI = 95% confidence interval.
***Medical history is a count of the following conditions: chronic hypertension, renal disease, diabetes millitus, and previous pre-eclampsia.**

Patients with at least one pre-existing medical condition were again found to have a lower risk of antenatal adverse events compared to those with no pre-existing medical conditions. As previously stated, this finding is consistent with existing prognostic models that are currently in use (von Dadelszen et al., 2009, Thangaratinam et al., 2017). The reduction in risk was also present in the cause-specific model for delivery. A potential explanation for this association may be that patients with additional complications were more likely to seek or receive pre-natal care, which is known to reduce the risks of complications and early delivery (von Dadelszen et al., 2009). Both antihypertensive and magnesium sulphate treatments were found to significantly increase the risk of delivery, and magnesium sulphate treatment was again found to increase the risk of antenatal events, in the cause-specific models. This is expected, as these treatments are given to high risk patients to minimise the risk of eclampsia its associated complications.

The two competing risks approaches are applied to estimate the cause-specific cumulative incidence function for antenatal adverse events. However, the model selection procedure resulted in a simpler model being selected for the subdistribution approach (containing five prognostic factors) and a more complex model for the cause-specific approach (containing all nine prognostic factors for delivery and eight for antenatal adverse events). The subdistribution model is more parsimonious, however further investigations are required to assess the prognostic ability of the models.

5.3.4 Comparison of baseline cumulative incidence functions

The risk of experiencing an antenatal adverse event over time for a participant with average continuous prognostic factor, and zero categorical prognostic factor values is estimated as the baseline cause-specific cumulative incidence. Estimates of the baseline cumulative incidence function for antenatal adverse events from the

subdistribution and cause-specific models are reported in Table 4.6, at given time points, and over time in Figure 5.6.

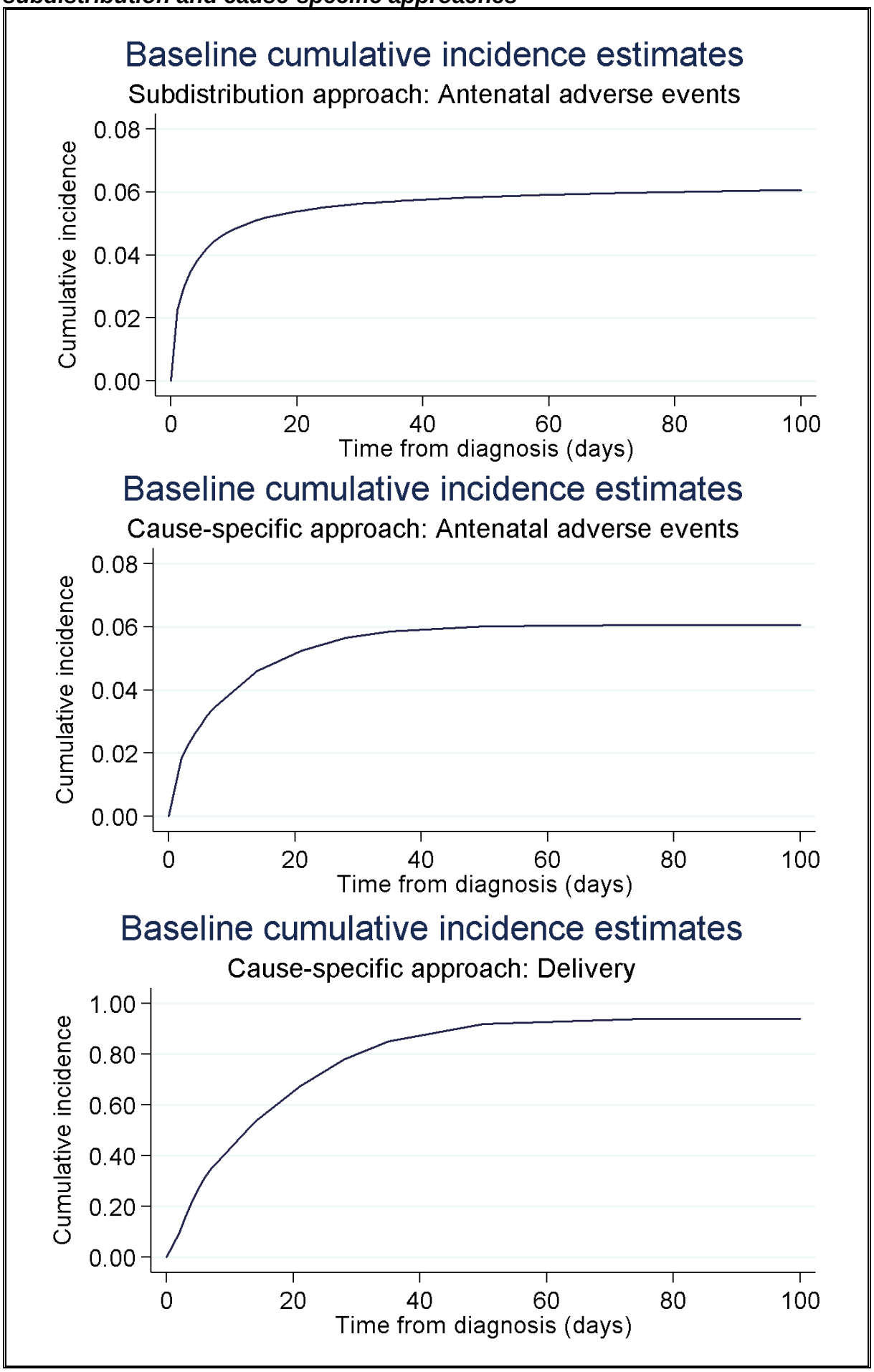
Table 5.5: Baseline cumulative incidence estimates for antenatal adverse events from fitted competing risks prognostic models

Time	2 days	1 week	4 weeks
Subdistribution model	0.030	0.044	0.056
Cause-specific model	0.018	0.034	0.056

The baseline cumulative incidence estimates from each of the competing risks approaches indicate the risk of antenatal adverse events increases steeply in the first three weeks after pre-eclampsia diagnosis. After this time the absolute risk remains stable at around 6%. The subdistribution model predicts a steeper increase in risk in the first week after pre-eclampsia diagnosis compared to the cause-specific model. However, the difference in absolute risks at this time is 1%, and the predictions made by the models tend to agree at later time points (three weeks after diagnosis).

The cause-specific approach requires estimates from all events to be combined, thus incorporates baseline cumulative incidence estimates for delivery (also depicted in Figure 5.6). The predicted risk of delivery increases steadily the first five weeks following pre-eclampsia diagnosis, on a much larger scale to antenatal adverse events (note the y axis). The absolute risk of delivery stabilizes at 94%, as deliveries are more common than antenatal adverse events.

Figure 5.6: Baseline cumulative incidence functions estimated using subdistribution and cause-specific approaches



5.3.5 Sensitivity analysis

A sensitivity analysis was conducted to test the assumptions made during model development. The proportional hazards assumption, made by both the cause-specific and subdistribution approaches, was tested by incorporating interactions between the included prognostic factors and log time, using a linear (one degree of freedom) spline function. The results are given in Appendix X; there is evidence that the effect of magnesium sulphate treatment was not constant over time in both the subdistribution and cause-specific approaches ($p<0.001$), and thus the proportional hazards assumption was violated for all fitted models. Additionally, the proportional cause-specific hazards assumption was violated for systolic blood pressure ($p=0.003$) and maternal age ($p=0.005$) for the delivery model. Again, these non-proportional estimates were intentionally disregarded due to the small sample size of this study and the concern for overfitting. Again the inclusion of significant time interactions was thought unlikely to affect the prognostic performance of the models, as the time points assessed occur soon after diagnosis (2 days, 1 week, and 4 weeks). In general, the assumption of proportional subdistribution hazards does not hold if the assumption of proportional cause-specific hazards is true (Beyersmann et al., 2009). The results from the subdistribution model may be interpreted as time-averaged effects on the cumulative incidence scale, which may affect the interpretation of the model predictions. However, not incorporating these interactions is unlikely to affect the results of this chapter, which focuses on the comparison of the two competing risks modelling approaches.

5.3.6 Comparison of competing risks prognostic model predictions

The flexible parametric competing risks prognostic models were applied to the PREP study participants to predict a participant's risk of antenatal adverse events, accounting for the presence of competing events. Although these are not the final models, comparisons at this stage are presented as the subdistribution hazard will be

adjusted for optimism while the cause-specific hazard model will not. To enable predictions for the full set of individuals used within model development, any missing information was assigned the median imputed value for each missing prognostic factor for each participant. The median cause-specific cumulative incidence estimates for antenatal adverse events for both modelling approaches are reported in Table 5.6.

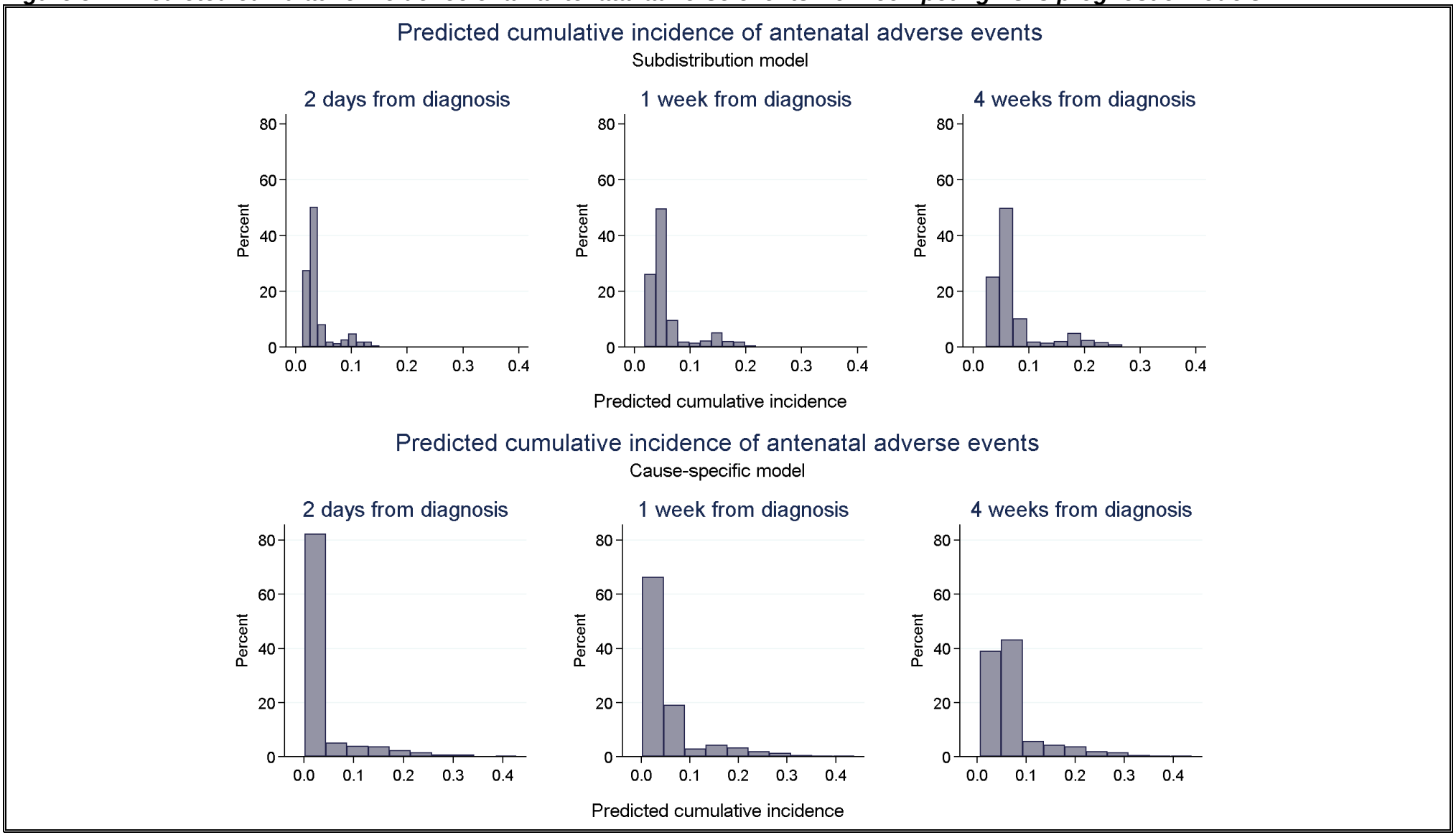
Table 5.6: Median cumulative incidence estimates for antenatal adverse events

	2 days	**1 week**	**4weeks**
Subdistribution model	3.2%	4.8%	6.0%
Cause-specific model	1.9%	3.5%	5.7%

The average predicted risk of experiencing an antenatal adverse event estimated using the cause-specific model is lower at all time points compared to those from the subdistribution model. However, the differences are small (1.3% at two days and one week) and lessen as time progresses (0.3% at four weeks). Distributions of the predicted cumulative risk of antenatal adverse events for PREP study participants at two days, one week, and four weeks estimated using the two competing risks prognostic models are displayed in Figure 5.7.

Both models produced right-skewed risk prediction distributions, with the majority of participants predicted to have a low risk (<10%) of experiencing an antenatal adverse event. For both models the distribution of predicted risks shift steadily upwards as time progresses. The range of predicted risks estimated using the subdistribution model grows over time, with a maximum predicted risk of 15.0% at two days, and 26.8% at four weeks. Whereas the range of predicted risks estimated using the case-specific model remains stable, with maximum values of 42.8% and 43.8% at two days and four weeks respectively.

Figure 5.7: Predicted cumulative incidence of an antenatal adverse events from competing risks prognostic models



5.3.7 Apparent prognostic performance of the fitted prognostic models

The median predicted cumulative risk of experiencing an antenatal adverse event at two days, one week, and four weeks for each of the four risk groups are presented in Table 5.7. The average predicted risks estimated using the subdistribution model are greater than the average risks estimated using the cause-specific model at all time points for all but the highest risk group.

Table 5.7: Median predicted cumulative incidence for risk groups at given time points

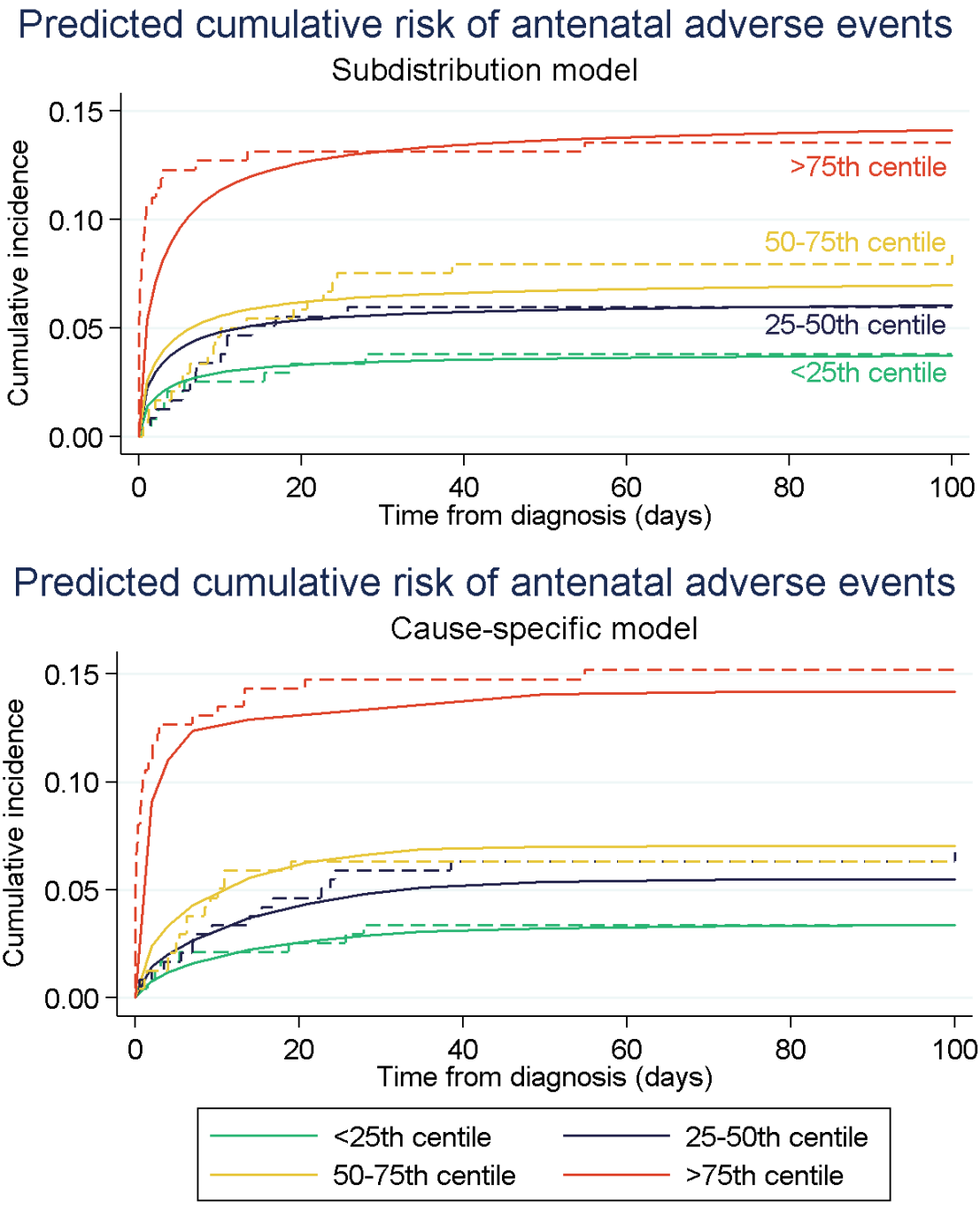
	Subdistribution model			Cause-specific model		
	2 days	**1 week**	**4weeks**	**2 days**	**1 week**	**4weeks**
<25th centile	1.8%	2.7%	3.4%	0.6%	1.6%	2.9%
25-50th centile	3.0%	4.4%	5.6%	1.4%	2.7%	4.8%
50-75th centile	3.4%	5.1%	6.4%	2.4%	4.3%	6.6%
>75th centile	7.1%	10.5%	13.1%	9.1%	12.4%	12.9%

The average predicted cumulative incidence of antenatal adverse events for the four risk groups over time are presented in Figure 5.8 (solid lines) for each model. These are compared to observed risks (dashed lines), calculated using the non-parametric cumulative incidence functions which account for competing events.

The predicted cumulative risks of both models struggle to capture the steep increase in cumulative incidences directly after pre-eclampsia diagnosis. The steep increase in risk observed in the highest risk group (>75th centile: red lines) directly after pre-eclampsia diagnosis is better captured by the cause-specific model than the subdistribution model. The average risk predictions from the subdistribution model agree with the observed risks from three weeks following diagnosis for all groups apart from the second highest (50-75th centile: yellow line). The average predicted risks from the cause-specific model agree with the observed risks up to three weeks following diagnosis for all risk groups, and at all times for the lowest risk group (<25th centile:

green line). There is little observable difference between the average risk in the middle two risk groups (25-50th centile: blue line, 50-75th centile: yellow line) after 40 days, though the cause-specific model predicts some difference between the groups.

Figure 5.8: Predicted cumulative incidence of antenatal adverse events by risk groups for competing risks prognostic models



The solid lines represent the average predicted cumulative incidence of antenatal adverse events for the four risk groups over time.
The dashed lines represent average observed risks for each risk group, calculated using non-parametric cumulative incidence functions which account for competing events.

The expected and observed risks of antenatal adverse events at two days, one week, and four weeks were calculated for the PREP study participants using both prognostic models. These measures are compared in Table 5.8. The expected risks from the subdistribution model are closer to the observed risks than those of the cause-specific model at all time points. The overall calibration performance of the subdistribution model is good (E/O>0.9) two days and one week after diagnosis, the performance decreases for the last time point. The overall calibration performance of the cause-specific model improves over time.

Table 5.8: Apparent overall calibration of competing risks prognostic models

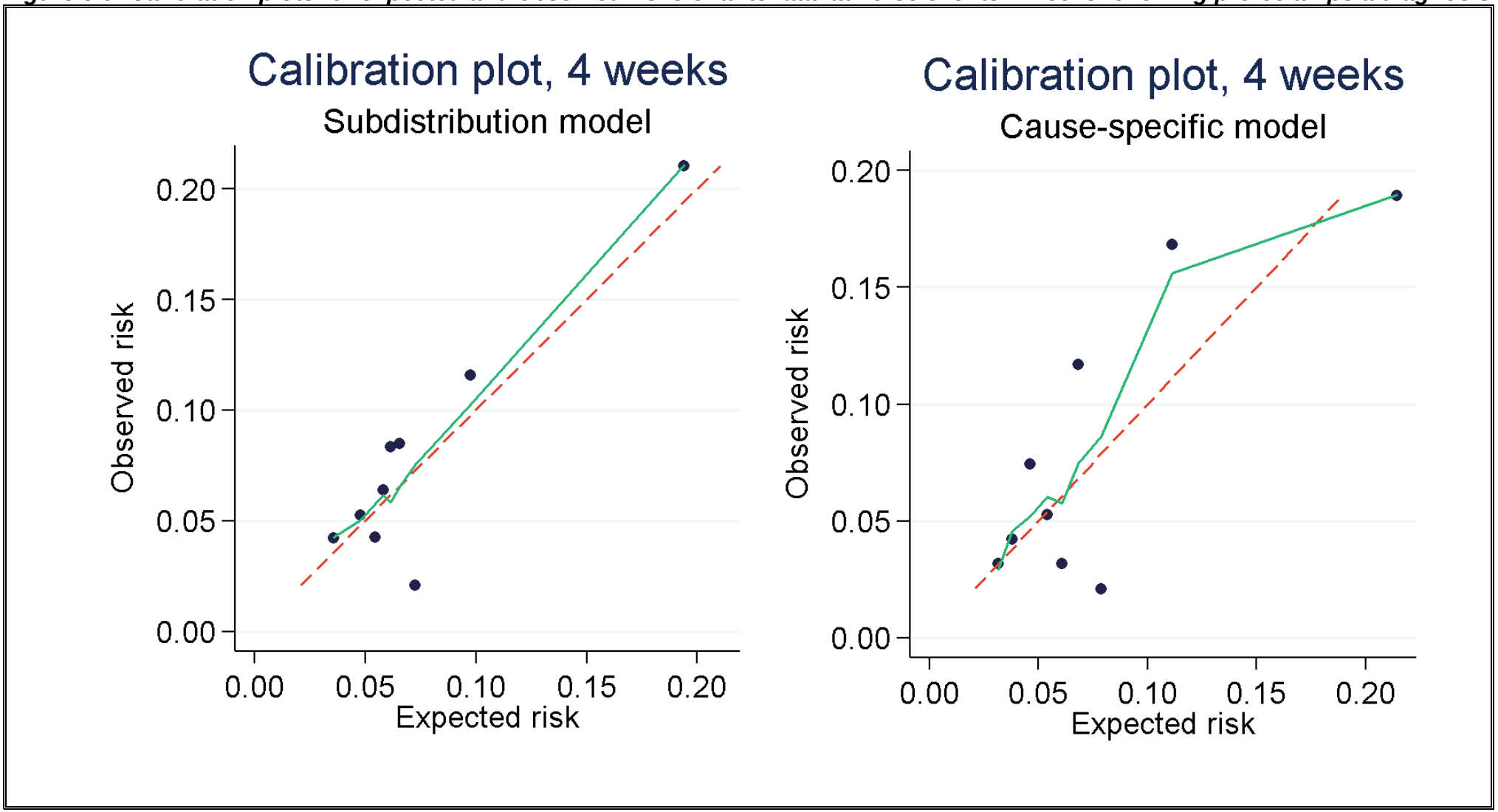
	Non-parametric CIF estimate	Subdistribution model		Cause-specific model	
		Expected	E/O	Expected	E/O
2 days	3.5%	3.2%	0.914	1.9%	0.543
1 week	5.3%	4.8%	0.906	3.5%	0.660
4 weeks	7.6%	6.0%	0.789	5.7%	0.750

The apparent calibration slope for the subdistribution model was 1.001[i] (95% CI: 0.65, 1.35) in the data in which the model was developed, as expected. Calibration plots of expected and observed risks at four weeks for both models are depicted in Figure 5.9. The subdistribution model again appears to perform slightly better than the cause-specific model.

A potential explanation for the difference in calibration performance between the two models may be due to the requirement of the cause-specific model to estimate parameters for both the event of interest and the competing event. This approach may thus require a larger sample size to accurately estimate the cumulative incidence function for an event, hence resulting in poorer calibration performance.

[i]The apparent calibration slope was found not to be exactly 1 because of the multiple imputation methods used in both model development and validation.

Figure 5.9: Calibration plots for expected and observed risks of antenatal adverse events 4 weeks following pre-eclampsia diagnosis



Discrimination was assessed using Harrell's C-index at two days, one week, four weeks, and overall, the D-statistic, which measures prognostic separation, and R^2_D (the proportion of explained variation). The resulting measures of discrimination are given in Table 5.9. Both models are better able to discriminate between those who experience the outcome of interest and those who do not two days after diagnosis than four weeks after diagnosis. The cause-specific model produces larger C-index measures at all time points. The subdistribution model explains 16.5% of the variation between the two prognostic outcomes, whereas the cause-specific model (ranked using 4 week predictions[i]) explains 15.2% of the variation.

Table 5.9: Harrell's C-index, D-statistic, and R^2_D for discrimination for fitted prognostic models (apparent performance)

	Harrell's C-index				Royston and Sauerbrei's D-statistic	
	2 days	**1 week**	**4 weeks**	**Overall**	**D-statistic**	**R^2_D**
Subdistribution model	0.824	0.721	0.649	0.652 (0.59, 0.72)	0.908	0.165
Cause-specific model	0.836	0.753	0.678	[ii]	0.868	0.152

In summary, the apparent predictive performance of the competing risks prognostic models differ slightly. The subdistribution model is well calibrated and better able to predict the observed risks of antenatal adverse events, whereas the cause-specific model is better able to discriminate between those who will and won't go on to experience an antenatal adverse event. Both model's ability to discriminate between the two competing events lessen as time progresses, as expected.

5.3.8 Internal validation and shrinkage of subdistribution prognostic model

[i]The D-statistic uses the ranked order of the predicted risks of a model to assess prognostic separation. However as the cause-specific model does not have a 1:1 relationship between the risk and rate, the ranked order of the predictions may change over time. Thus the D-statistic requires ranking to be assessed at a specified time-point.

[ii]Similarly, Harrell's C-index cannot be calculated for the cause-specific model over all time points due to loss of 1:1 relationship between hazards and risk.

The model developed using the cause-specific approach was not internally validated due to the time consuming, computationally intensive nature of the simulation approach for obtaining individual risk predictions from this model. The optimism in performance of the subdistribution model was assessed using the bootstrap procedure outlined in Box 4.2. The bootstrap performance measures reflect the model's performance in the imputed bootstrap data in which it was developed. The test performance measures reflect the model's performance when assessed using the original PREP study data. The optimism is calculated as the difference between the bootstrap and the test measures. The average model performance measures are reported below in Table 5.10.

Table 5.10: Internal validation measures of prognostic performance for subdistribution prognostic model in 100 bootstrap samples

	Subdistribution model			
	Average bootstrap performance	**Average test performance**	**Average optimism**	**Optimism-adjusted performance**
Calibration Slope	1.000	0.760	0.240	0.760
Harrell's C-index	0.694	0.647	0.046	0.648
D Statistic	1.156	0.889	0.281	0.875
R2D Statistic	0.242	0.159	0.085	0.157

The average optimism in the measures of discrimination, Harrell's C-index and R^2_D, for the subdistribution model were 0.046 and 0.085 respectively. Subtracting the average optimism from the fitted model's apparent performance (see Table 5.9), results in an optimism-adjusted C-index of 0.652-0.046=0.606 and an optimism-adjusted R^2_D of 0.165-0.085=0.080. The average optimism observed in the subdistribution hazards model calibration slope was 0.240, resulting in a uniform shrinkage factor of 1.001-0.240=0.760.

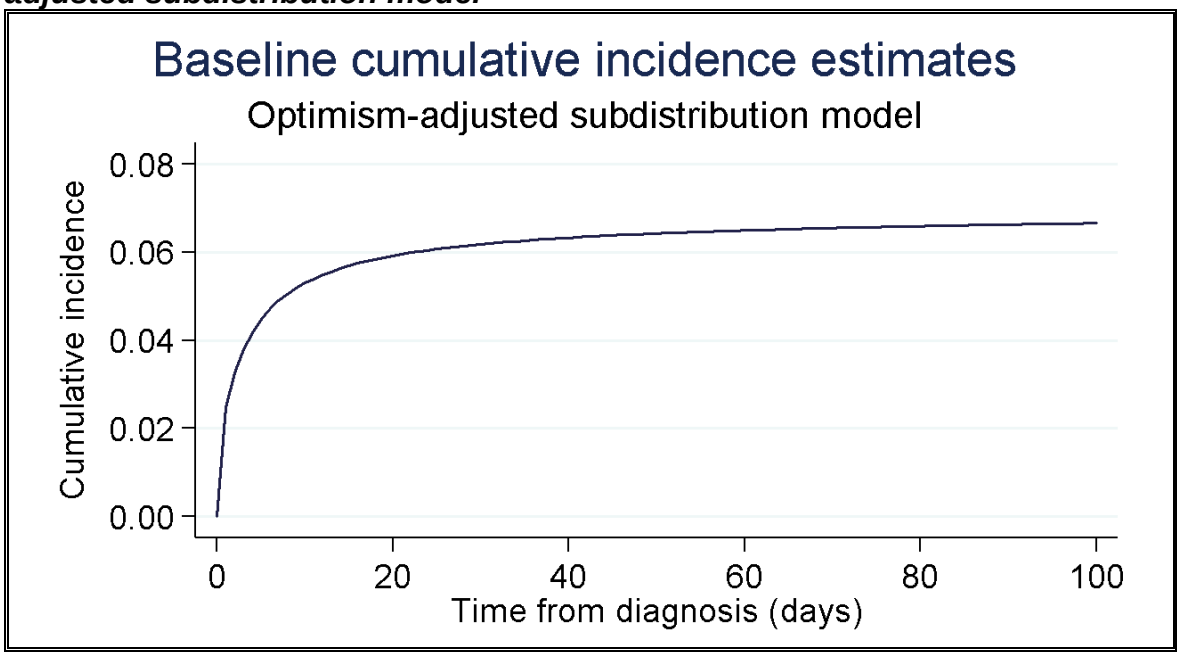
To adjust for the overfitting (optimism), the shrinkage factor for the subdistribution model (i.e. the optimism-adjusted calibration slope) was applied uniformly to all predictor coefficients within the prognostic model, and the baseline cumulative incidence function was re-estimated to ensure calibration-in-the-large. The final optimism-adjusted regression equation for the subdistribution model is provided in Table 5.11, and the updated baseline cumulative incidence function for the optimism-adjusted model is presented in Figure 5.10.

Table 5.11: Final regression equation for optimism-adjusted subdistribution model

Optimism-adjusted subdistribution model	
$LP = -\ 0.005 \times (\boldsymbol{maternal\ age} - 30.245) - 0.038 \times (\boldsymbol{gestational\ age} - 30.562) - \begin{cases} 0.474, & if\ \boldsymbol{one\ preexisting\ condition} \\ 0.142, & if\ \boldsymbol{two\ or\ more\ conditions} \end{cases} + 0.135 \times (if\ \boldsymbol{antihypertensive\ treatment}) + 0.869 \times (if\ \boldsymbol{MgSO4\ treatment})$	$F_k(t\|\mathbf{0}) = \begin{cases} 0.033, & if\ t = 2\ days \\ 0.049, & if\ t = 1\ week \\ 0.061, & if\ t = 4\ weeks \end{cases}$
$F_k(t\|\boldsymbol{X}_k) = 1 - \left(1 - F_k(t\|\mathbf{0})\right)^{exp(LP)}$	
$Pr(Antenatal\ adverse\ event) = 1 - S_0(t)^{exp(\hat{\boldsymbol{\beta}}\boldsymbol{X})}$	

$\mathbf{LP}$ = Linear predictor
$\mathbf{F_k(t|0)}$ = Baseline cause-specific cumulative incidence function
$\mathbf{F_k(t|X_k)}$ = Cause-sepcific cumulative incidence function

Figure 5.10: Baseline cumulative incidence function estimated using optimism-adjusted subdistribution model



5.4 Discussion

In this chapter, two approaches for analysing time-to-event data in the presence of competing risks were applied and compared. Two flexible parametric competing risk models were developed to predict the risk of antenatal adverse events in patients diagnosed with pre-eclampsia, accounting for delivery of the baby as a competing event. The key findings and conclusions are summarised below.

5.4.1 Key clinical findings

The two prognostic models developed in this chapter can be applied to patients diagnosed with early-onset pre-eclampsia to predict their future risk of antenatal adverse events, in the real-world setting where the baby may be delivered before any antenatal adverse events have occurred. These models build on the work of Thangaratinam et al who developed models to predict the risk of the composite outcome of adverse events or early delivery in early-onset pre-eclampsia patients (Thangaratinam et al., 2017). The competing risks analyses applied in this chapter facilitates the disentanglement of these two outcomes, thus aiding the understanding of a patient's real world risks of experiencing adverse events prior to spontaneous or induced labour.

Antenatal adverse events were experienced by 7.9% of the participants in the PREP study cohort, with a cumulative incidence at 100 days of 7.8%. It is shown that when assessing the cumulative incidence of these events, not accounting for delivery as a competing event results in a biased estimate nearly twice the size of the true value (Figure 5.2).

5.4.2 Key statistical findings

Two approaches were utilised to account for competing events during prognostic model development, the cause-specific and subdistribution approaches. Both aim to

predict the same function; however, the cause-specific cumulative incidence function produced by these methods are noticeably different. A number of statistical advantages were identified for using the subdistribution approach over the cause-specific approach in prognostic model research.

Firstly, the subdistribution approach estimates the direct effect of the included prognostic factors on the real world risk of the event of interest through subdistribution hazard functions. Subdistribution hazards adjust for the indirect effects of the prognostic factors on the competing events. The cause-specific approach estimates the direct effect of the included prognostic factors on the event of interest, while ignoring (censoring) the competing event. Cause-specific hazards are thus better suited to prognostic factor research than prognostic modelling.

Secondly, the subdistribution approach retains the one-to-one relationship with the cause-specific cumulative incidence function. This aids the interpretation of subdistribution hazard ratios as an increasing hazard ratio indicates an increased cumulative incidence, which is not always the case with cause-specific hazard ratios. The one-to-one relationship also simplifies estimation of the cumulative incidence function; the subdistribution approach does not require the estimation of hazard functions for all outcomes, whereas the cause-specific approach does. The added complexity of the cause-specific model requires more parameters to be estimated for the model, resulting in the need for a larger sample size to develop a prognostic model.

Finally, the subdistribution approach results in a model of a similar form to that of the standard time-to-event analysis approach. Obtaining individual predictions from these models is simple, as is adapting standard modelling processes (such as shrinkage) for competing events, and comparisons between these and standard time-to event models is intuitive. The complexity of the cause-specific approach means that additional measures, such as simulation methods, are require to obtain individual

predictions, there is some difficulty in adapting internal validation processes, and comparisons with standard time-to-event models is difficult.

Overall, the subdistribution approach is better suited to prognostic model research, it often results in simpler models which are easier to understand, apply, and internally validate, compared to the cause-specific approach.

5.4.3 Strengths and limitations

In the current literature, relatively few prognostic models have been developed using competing risks statistical methods, the majority of which use semi-parametric modelling approaches to obtain hazard ratios without the need to estimate the baseline hazard functions (Mazzaferro et al., 2018, Scrutinio et al., 2018, Wolbers et al., 2009). The Fine and Gray (Fine and Gray, 1999) subdistribution approach features in the majority of the published articles, whereas comparison of models developed using this approach to a cause-specific model is uncommon (Wolbers et al., 2009). In this chapter, flexible parametric competing risks prognostic model were developed using the two modelling approaches. This allows for a comparison of the cause-specific cumulative incidence estimates, the primary purpose of prognostic model research, from both the subdistribution and cause-specific approaches.

Utilising competing risks statistical methods for the development of prognostic models is relatively novel, thus some of the methods required for adequate evaluation of prognostic models are yet to be developed. Little guidance is currently available for methods of multiple imputation or for the application of some internal validation methods, such as shrinkage, in the presence of competing events. Though convenient solutions were found to enable the analysis to proceed for the purpose of this thesis, additional research on the true effects of the methods employed and any biases produced is required to establish if the findings are robust.

5.4.4 Further work

The prognostic models developed in this chapter highlight the importance of incorporating the competing risks of delivery when assessing patient risks during pregnancy. A comparison of pregnancy models developed using competing risks and standard time-to-event methods will further highlight the differences in predicted risks, and the need to account for competing events to ensure accuracy of real world predictions in prognostic models. An external validation study of the prognostic models developed within this thesis for the prediction of antenatal adverse events will enable such a comparison. The following chapter will directly compare and externally validate the models developed in Chapters 4 and 5.

6 EXTERNAL VALIDATION OF PROGNOSTIC MODELS DEVELOPED USING STANDARD AND COMPETING RISKS METHODS

6.1 Introduction

Prognostic models to predict the risk of antenatal adverse events in women diagnosed with early-onset pre-eclampsia were developed in previous chapters (4 & 5). In this chapter, the key models are externally validated.

In Chapter 4, the two prognostic models developed using standard time-to-event (Cox and Royston-Parmar) approaches, which did not appropriately account for the competing events, performed similarly during internal validation. However, the Royston-Parmar model was developed using a flexible parametric approach which enabled the direct estimation of a smooth underlying baseline hazard function. Therefore, the optimism adjusted Royston-Parmar model (and not the Cox model) is externally validated in this chapter. This model will henceforth be referred to as the *"PREP-RP"* model.

Similarly, in Chapter 5 two prognostic models were developed using flexible parametric competing risks (subdistribution and cause-specific approaches) approaches. The subdistribution model performed better than the cause-specific model during apparent validation, and a number of other statistical advantages of this model were identified. For these reasons, the optimism-adjusted subdistribution model (and not the cause-specific model) is externally validated in this chapter. This model will henceforth be referred to as the *"PREP-SD"* model.

6.1.1 Overview of validation methods for competing risks setting

It is recommended to evaluate the performance of a prognostic model using data independent from those used to develop the model (Collins et al., 2016), via external validation; especially where the generalisability of the model needs to be assessed for a wider range of populations or settings. It is also important to account for competing events, if present, when validating a time-to-event prognostic model. The assessment of model calibration compares absolute risk predictions from the prognostic models to estimates of observed risk (Royston and Altman, 2013). If the methods used to estimate the observed risk do not appropriately account for the competing events (for example, using risk estimates derived via the Kaplan-Meier rather than the actual cumulative incidence function) then the resulting model calibration measures will be biased (Wolbers et al., 2009). The subdistribution hazard risk set, which retains participants in the risk set following the occurrence of a competing event, is used to appropriately account for competing events when calculating the cumulative incidence function during external validation.

The *PREP-RP* model, developed in Chapter 4, was developed using a standard time-to-event analysis approach in which deliveries were censored as they occurred. However, the model only utilises information from the event of interest (antenatal adverse events) and does not incorporate the hazards for the competing event (deliveries) and so does not appropriately manage the competing events. Therefore, this model is expected to perform poorly during external validation, particularly when measuring calibration, as it is likely to overestimate the risk of antenatal adverse events in the presence of competing events. Whereas the *PREP-SD* model, developed in Chapter 5, incorporates the risks of the competing event and so is expected to produce representative estimates of absolute risks, and thus perform better than the *PREP-RP* model in terms of measures of calibration during external validation.

6.1.2 External validation dataset

Both the *PREP-RP* and *PREP-SD* models were developed using data from the PREP study (Thangaratinam et al., 2017). The original prognostic models developed in this study ("PREP-L" and "PREP-S") were externally validated using data from the Pre-eclampsia Integrated Estimate of RiSk for mothers (PIERS) cohort (von Dadelszen et al., 2009). The aim of the PIERS model research programme was to develop an evidence based framework to enable clinicians to better define maternal risks associated with diagnosis of pre-eclampsia (von Dadelszen et al., 2009). The study cohort included 1,259 pregnant women admitted to hospital with pre-eclampsia, diagnosed at any time during pregnancy. The study collected baseline participant information for similar prognostic factors as examined in the PREP study, as well as time-to-event information for comparable adverse event outcomes.

6.1.3 Aims

The overall aim of this chapter is to investigate the impact of the management of competing events during prognostic model development on measures of predictive performance. Data from the PIERS study cohort was used to externally validate the *PREP-RP* and *PREP-SD* prognostic models. The measures of calibration and discrimination reported in this chapter are adapted to appropriately account for competing events. Differences between the two model's predictions and prognostic performance are discussed.

6.2 Methods: External validation of prognostic models

Methods for the external validation of the *PREP-RP* and *PREP-SD* models using data from the PIERS study cohort are now discussed. The following methods are structured to adhere to the TRIPOD guidelines (Collins et al., 2015) for reporting prediction model development and validation studies.

6.2.1 External validation study participants

The PIERS study cohort includes participants who were admitted to hospital following a diagnosis of pre-eclampsia, and participants diagnosed with pre-eclampsia following delivery of the baby. The PREP study (and thus the *PREP-RP* and *PREP-SD* models) focuses on participants with early-onset pre-eclampsia (diagnosed prior to 34 weeks gestation) and events which occur after diagnosis. To create a comparable external validation dataset, a subset of the PIERS study participants was selected. This subset, henceforth be referred to as the "*validation set*", contains PIERS cohort participants with:

1. a confirmed diagnosis of pre-eclampsia prior to 34 weeks gestation, and
2. a diagnosis of pre-eclampsia prior to adverse events or delivery of the baby, and
3. a known delivery date, and
4. no missing information for all prognostic factors contained in the *PREP-RP* and *PREP-SD* models.

A flow diagram depicting the exclusion of PIERS participants to form the validation set is provided in the results.

6.2.2 Outcome definition

The definition of antenatal adverse events differs from that outlined in Section 4.2.1, due to some disparities in the recording of the component events in the PIERS study. Although the majority of adverse events investigated in the PREP study were

also investigated in the PIERS study, some differences exist. Information on the following adverse events were not recorded in the PIERS study: retinal detachment, posterior reversible encephalopathy, subcapsular haematoma, postpartum haemorrhages, and placental abruptions. As none of the PREP study participants experienced a retinal detachment, posterior reversible encephalopathy, or subcapsular haematoma as their first event, and as postpartum haemorrhages were excluded from the definition of *antenatal* adverse events (as they occur after delivery of the baby), the absence of information on these adverse events is unlikely to impact the validation performance of the prognostic models. However, placental abruptions were the most commonly observed antenatal adverse events in PREP study participants (1/3 of all antenatal adverse events). The absence of information on this adverse event is likely to affect the validation performance of the prognostic models in these data. A summary of the incidence of antenatal adverse events in the PIERS validation set is reported and compared to the PREP study; the sample size is discussed.

The time to antenatal adverse events in participants with early-onset pre-eclampsia is defined in Section 5.2.1. Participants are at risk from the date of confirmed pre-eclampsia diagnosis, and remain at risk until they experience either an antenatal adverse events or delivery the baby. Unlike the PREP study, the PIERS study data only included the date (and not the time) of a number of outcomes, so for simplicity this analysis only included date information. A random follow-up time (taken from a Uniform[0, 0.01] distribution) was generated for participants who experienced an event on the same date as pre-eclampsia diagnosis, to retain the participant in the analysis. If both an antenatal adverse event and delivery of the baby occurred on the same day, the outcome was recorded as an antenatal adverse event (i.e. assuming the adverse event occurred just before delivery). This assumption was previously used in the PREP dataset, and reflects “real world” clinical practice in which an adverse event prompts induction of labour.

6.2.3 Descriptive analysis of external validation (PIERS) participants

The average follow-up time (calculated using the reverse Kaplan-Meier estimator (Schemper and Smith, 1996)) and non-parametric cumulative incidence estimates of pre-eclampsia events for the validation participants are reported and compared to the model development population.

A descriptive analysis of prognostic factors included in the *PREP-RP* and *PREP-SD* models in the validation set was performed. The distribution of each continuous prognostic factor was investigated; the mean and standard deviation are reported for prognostic factors which were normally distributed, whereas the median and interquartile range are reported for those with skewed distributions. The frequency and percent of participants in each category of binary and categorical prognostic factors are also reported. The resulting prognostic factor summaries are compared to those from the model development data to investigate the relatedness of the populations (Debray et al., 2015).

Due to the small number of participants missing prognostic factor and outcome information, multiple imputation methods were not considered necessary. Thus, a complete case analysis was performed in this validation study.

6.2.4 Deriving predicted risks of antenatal adverse events

Individual linear predictor and cumulative risk estimates were obtained using the equations reported in Table 4.11 (*PREP-RP* model) and Table 5.11 (*PREP-SD* model). For each model, linear predictor estimates for each participant in the validation set were obtained by applying the optimism adjusted regression coefficients and prognostic factor transformations to the participant data. The distributions of the linear predictor estimates in the validation set for the *PREP-RP* and *PREP-SD* models are compared to linear predictor distributions in the model development population (Royston and Altman, 2013). For each model, individual estimates of the cumulative

risk of experiencing an antenatal adverse event two days, one week, and four weeks after early-onset pre-eclampsia diagnosis were estimated for each participant in the validation set. Individual risk estimates from the *PREP-RP* and *PREP-SD* models were compared at the given time points, as were the distributions of the predicted risks from each model for participants in the validation set.

6.2.5 External validation of prognostic models

The predictive performance of the *PREP-RP* and *PREP-SD* models was assessed in the validation set using measures of calibration and discrimination. For each model, four risk groups were created by splitting the validation set into equal sized groups using the 25th, 50th, and 75th centiles of the model's cumulative risk predictions at four weeks[i]. Overall calibration was examined by comparing the expected event risk (calculated using cumulative risk estimates from *PREP-RP* and *PREP-SD* models) to the observed event risk (estimated using non-parametric cumulative incidence function for the validation set). Calibration was assessed for each model at two days, one week, and four weeks following pre-eclampsia diagnosis using the validation set as a whole, and for each of the aforementioned risk groups. For each model, the average predicted risk in each risk group over time is presented alongside non-parametric cumulative incidence curves, enabling a comparison of the expected and observed risk over time.

For each model, the distribution of predicted risk for participants who had experienced an antenatal adverse event (event of interest) by four weeks was compared to the distribution of predicted risk for participants who had delivered a baby (competing event) by four weeks. The discriminative ability of the *PREP-RP* and *PREP-SD* models was assessed using Harrell's C-index (Harrell et al., 1982), the D-statistic, and R^2_D. The C-index was calculated as an overall measure and at two days,

[i]These risk groups created for each of the *PREP-RP* and *PREP-SD* models are likely to contain different sets of individuals.

one week, and four weeks, and the D-statistic and R^2_D was calculated and compared for both the *PREP-RP* and *PREP-SD* models.

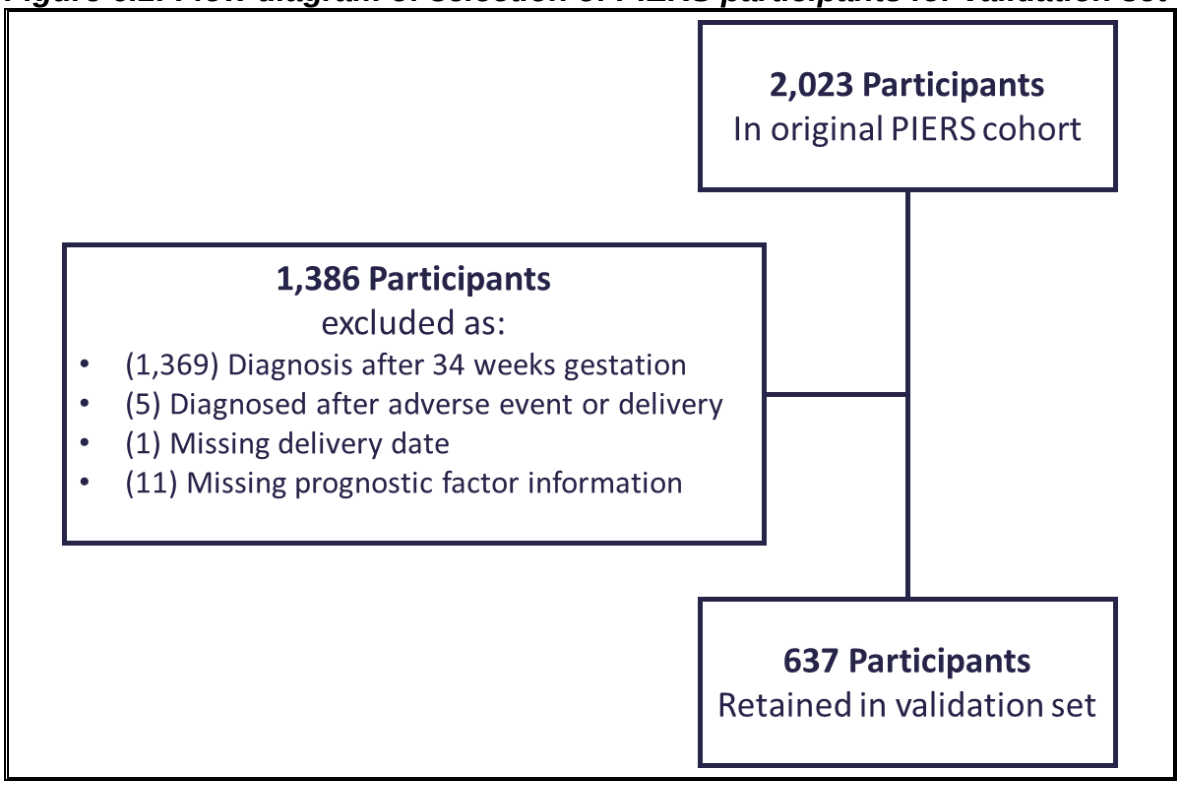
6.3 Results: External validation of prognostic models

The results of the external validation of the *PREP-RP* and *PREP-SD* models using data from the PIERS study cohort are now discussed.

6.3.1 External validation study participants

The original PIERS study cohort includes 2,023 participants. A subset of the PIERS cohort were retained for this external validation study by applying four inclusion criteria (outlined in 6.2.1). A flow diagram depicting the selection of PIERS participants for the validation set is depicted in Figure 6.1.

Figure 6.1: Flow diagram of selection of PIERS participants for validation set



The validation set consists of 637 women with an early-onset pre-eclampsia diagnosis. A large proportion (1,369, 67.7%) of the PIERS study participants were excluded from the validation set as they were diagnosed with pre-eclampsia after 34 weeks gestation (i.e. not *early-onset* pre-eclampsia diagnoses). These participants are

expected to have fewer serious complications than those diagnosed before 34 weeks gestation (Sibai, 2003). As both the *PREP-RP* and *PREP-SD* models were developed in populations with early-onset pre-eclampsia, not excluding these participants is likely to adversely affect the validation performance of the models. Few participants were excluded for diagnosis after an adverse event or delivery (5, 0.2%) and for missing outcome or prognostic factor information (12, 0.6%).

6.3.2 Outcome definition

Disparities in the recording of the component events in the PIERS study results in a modified definition of antenatal adverse events in this validation study. Similar to the PREP study, none of the participants were lost to follow up, and all either experienced an antenatal adverse event (79 women, 12.4%), or delivered the baby without an adverse event. The incidence of first antenatal adverse events observed in the development (PREP) and validation set (PIERS) are reported in Table 6.1.

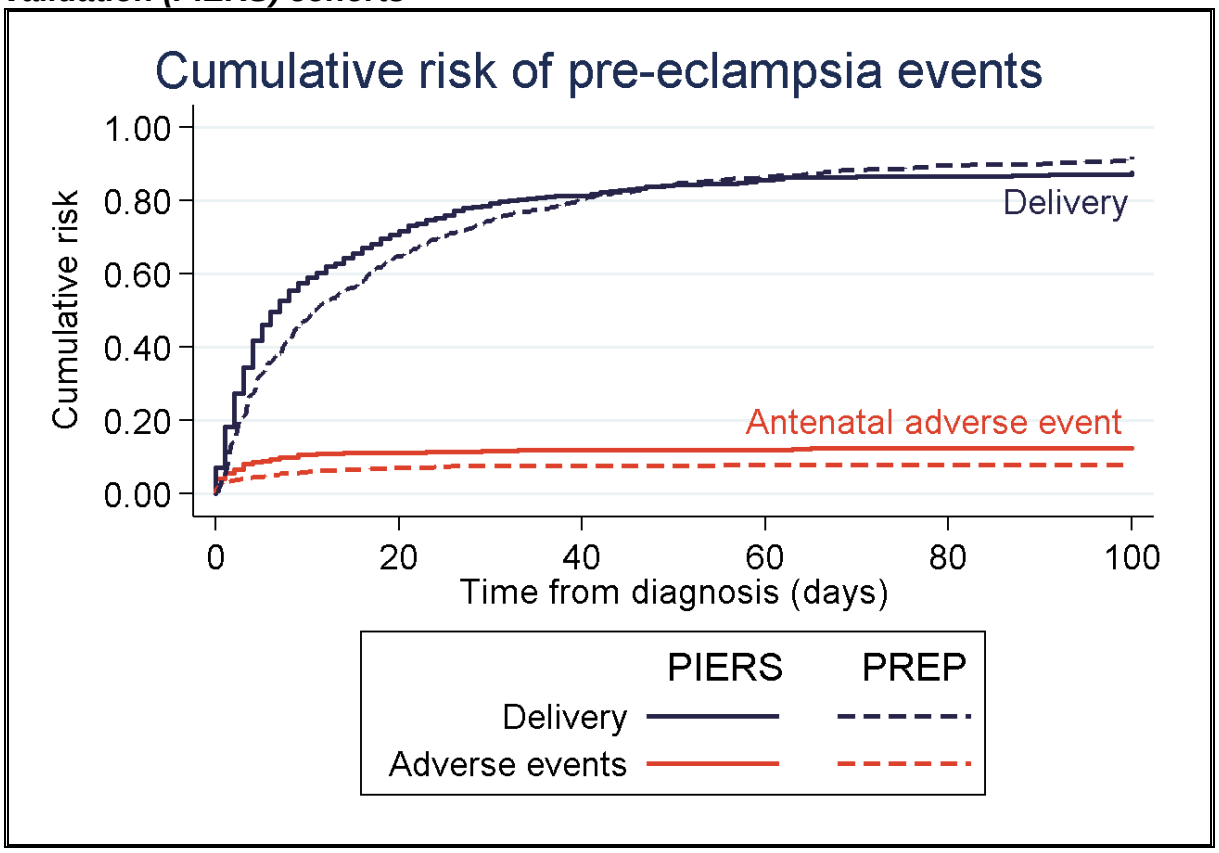
Table 6.1: Incidence of components of antenatal adverse events in development (PREP) and validation (PIERS) cohorts

First antenatal adverse event	PIERS cohort, Number (%)	PREP cohort, %
Pulmonary oedema	37 (5.8%)	0.4%
Transfusion of blood	19 (3.0%)	2.4%
At least 50% FIO2 for > 1hour	12 (2.0%)	0.1%
Acute renal insufficiency	4 (0.6%)	0.2%
Eclamptic seizure	3 (0.5%)	1.2%
Hepatic dysfunction	3 (0.5%)	0.5%
Positive inotrope support	1 (0.2%)	0
Total antenatal adverse events	**79 (12.4%)**	**7.9%**

A similar percentage of participants in both cohorts required a blood transfusion (3.0% in PIERS vs 2.4% in PREP) and experienced hepatic dysfunction (0.5% in both cohorts) as their first adverse event. A greater percentage of participants experienced a pulmonary oedema as a first adverse event in the PIERS cohort (5.8%) compared to the PREP cohort (0.4%), though the number of observed events is small. The overall proportion of all antenatal adverse events was greater in the PIERS cohort (12.4%) than the PREP cohort (7.9%). Although ideally an external validation sample requires a minimum of 100 events and 100 non-events for unbiased and precise estimation of predictive performance measures (Collins et al., 2016, Vergouwe et al., 2005), only 79 events are observed in this validation set. Regardless, this is the only cohort available to validate the two models without the need for a new cohort study which would take a number of years to conduct.

The participants in the validation set were followed up for a median of 7 days (IQR: 3 to 18), with the longest observed follow-up period equal to 114 days. The average follow-up is slightly shorter than that observed in the PREP study (median of 10.1 days). Non-parametric cumulative incidence estimates for the time to antenatal adverse events and delivery without an adverse event are depicted in Figure 6.2. A greater proportion of antenatal adverse events (red lines) were observed in the PIERS study (solid lines) than the PREP study (dashed lines).

Figure 6.2: Non-parametric cumulative incidence estimates of pre-eclampsia outcomes, accounting for competing risks, in development (PREP) and validation (PIERS) cohorts



6.3.3 Descriptive analysis of validation set

Results from the descriptive analysis of the prognostic factors included in the *PREP-RP* and *PREP-SD* models in both the development (PREP) and validation (PIERS) set are given in Table 6.2. The mean maternal age of participants at pre-eclampsia diagnosis is 31.2 years, with a median corresponding gestational age of 31.0 weeks (IQR: 28.4 to 32.9), similar to the PREP study participants. Just under half of participants did not have any historical medical conditions (44.4%), implying more PIERS study participants had at least one pre-existing medical condition (55.6%) in comparison to the PREP study (36.1%). The majority of participants were receiving antihypertensive treatment (87.0%) and just over half were receiving magnesium sulphate treatment (51.0%), more than observed in the PREP study participants (79.4% and 15.2% respectively).

Table 6.2: Descriptive analysis of prognostic factors in development (PREP) and validation (PIERS) cohorts

Prognostic Factor		PIERS Summary Statistics Mean(SD), N(%), Med [IQR]	PREP Summary Statistics Mean(SD), N(%), Med [IQR]
Maternal age (years)		31.2 (6.2)	30.2 (6.1)
Gestational age at diagnosis (weeks)*		31.0 [28.4 to 32.9]	31.4 [28.7 to 32.7]
Medical History¥	**0**	283 (44.4%)	594 (63.9%)
	1	251 (39.4%)	242 (26.0%)
	2 or more	103 (16.2%)	94 (10.1%)
Systolic blood pressure		168.3 (20.5)	158.6 (19.2)
Platelet count		205.0 (75.6)	227.1 (77.7)
Alanine amino transaminase*		25 [15 o 43]	17 [13 to 26.5]
Serum Creatinine		69.2 (20.4)	61.8 (17.1)
Antihypertensive treatment		554 (87.0%)	751 (79.4%)
Magnesium sulphate treatment		325 (51.0%)	144 (15.2%)

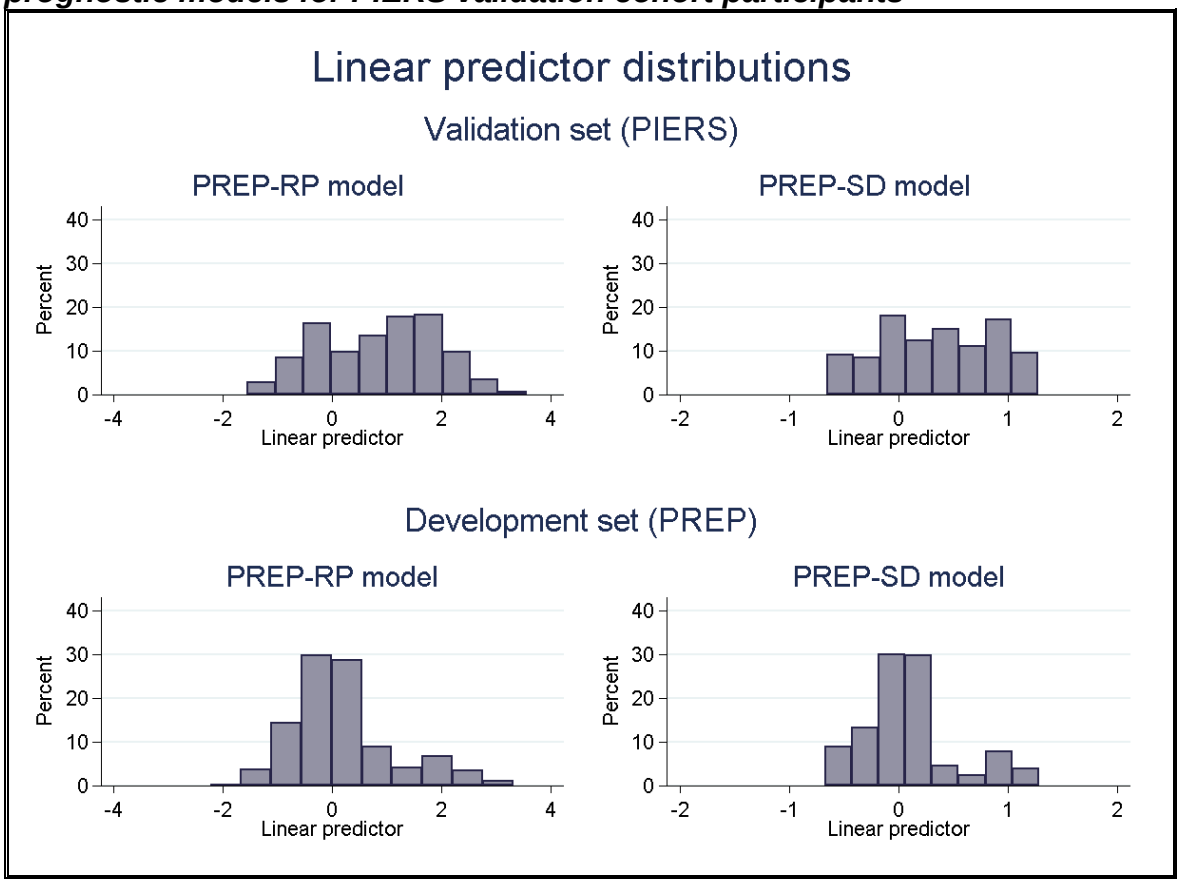
***Median and inter-quartile range (IQR) presented for non-normally distributed factors.**
¥Medical history is a count of the following conditions: chronic hypertension, renal disease, diabetes mellitus, and previous pre-eclampsia.

The difference in the number of patients with pre-existing medical conditions implies poorer health in the validation population, as demonstrated by higher SBP, ALT and serum creatinine measurements. In the developed models, patients with at least one pre-existing medical condition had reduce risks of antenatal adverse events, thus the poorer health of the population may explain the reduced incidence of the event of interest. The validation population also received more antihypertensive and magnesium sulphate treatments than the development population. This may be due to patients having poorer health, or could be an indicator of different care strategies for pre-eclampsia in Canada (validation study) and the UK (development study). In the developed models, patients receiving treatments had significantly increased risks of antenatal adverse events. The reduced incidence of antenatal adverse events in the validation study suggests the difference in proportion of patients treated may be due to different care pathways.

6.3.4 Deriving predicted risks of antenatal adverse events

The distributions of the individual linear predictor estimates for the *PREP-RP* and *PREP-SD* models in the development and validation set are provided in Figure 6.3. The range of linear predictor estimates for both the *PREP-RP* and *PREP-SD* models are similar across the development and validation sets. The shape of the distributions differs between studies; in particular, a greater proportion of the linear predictor estimates in the validation set are positive values, which may reflect the increase in risk of antenatal adverse events observed in the PIERS study (12.4%) in comparison to the PREP study (7.9%). Additionally, the linear predictor distributions observed in the validation set are more uniform than those observed in the development set, which clearly peak at estimates around zero.

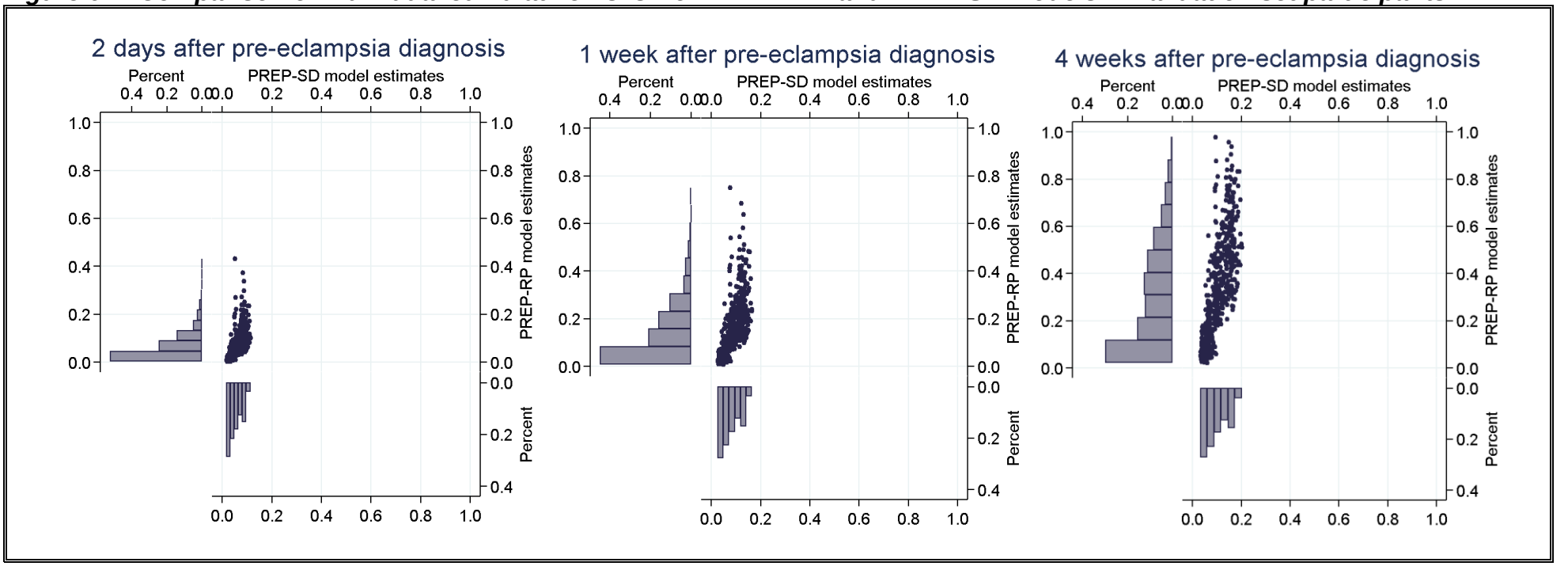
Figure 6.3: Distribution of individual linear predictor estimates from PREP prognostic models for PIERS validation cohort participants



The *PREP-RP* and *PREP-SD* model equations were applied to the validation set to predict individual cumulative risks of experiencing an antenatal adverse event by

two days, one week, and four weeks after pre-eclampsia diagnosis. The model predictions for each participant at each time point, and the distributions of the cumulative risks, are depicted in Figure 6.4. The range of individual risks estimated using the *PREP-RP* model, which censors deliveries, was larger than the *PREP-SD* model, which treats deliveries as competing events, for all time points. This is a result of the inflated absolute risk estimates produced by the *PREP-RP* model due to inappropriate management of competing events. The model predicts the hypothetical risk of experiencing an antenatal adverse event in participants who do not deliver without an adverse event. Whereas the *PREP-SD* model predicts “real-world” risks for all pre-eclampsia participants which account for the competing event. Thus, there is little observed agreement between individual risk estimates from the *PREP-RP* and *PREP-SD* models.

Figure 6.4: Comparison of individual cumulative risks from PREP-RP and PREP-SD models in validation set participants



6.3.5 External validation of prognostic models

The expected proportion of antenatal adverse events in the validation set by two days, one week, and four weeks after diagnosis was calculated using predicted cumulative risk estimates from each prognostic model. These expected outcome incidences were compared with the observed incidences of antenatal adverse events, accounting for the competing event of delivery, to provide a measure of overall calibration. The results are provided in Table 6.3. The overall expected risk of antenatal adverse events from the *PREP-RP* model exceeds the observed risk at later time points, as the model does not appropriately account for the competing events. Whereas the expected risks from the *PREP-SD* model underestimate the observed risks at these time points, which may be due to the differences in baseline risks of antenatal events (Figure 6.2).

Table 6.3: Overall calibration of prognostic models externally validated at specified time points

	Observed risk (O)	PREP-RP model		PREP-SD model	
		Expected risk (E)*	E/O	Expected risk (E)*	E/O
2 days	5.7%	5.6%	0.98	4.8%	0.84
1 week	9.4%	11.7%	1.24	7.1%	0.76
4 weeks	11.3%	26.3%	2.33	8.9%	0.79

***Expected risk calculated as median predicted risk for PIERS study participants derrived from prognostic model at time point**

Overall calibration was also assessed in four risk groups, created by splitting the validation set into equal sized groups using the 25th, 50th, and 75th centiles of each model's predicted cumulative risk at four weeks. The expected proportion of antenatal adverse event at two days, one week, and four weeks for each risk group are given in Table 6.4. The risk groups for the *PREP-RP* and *PREP-SD* model contain different sets of participants. In general, the predicted risks from the *PREP-RP* model are greater than those predicted by the *PREP-SD* model, with the exception of the low risk

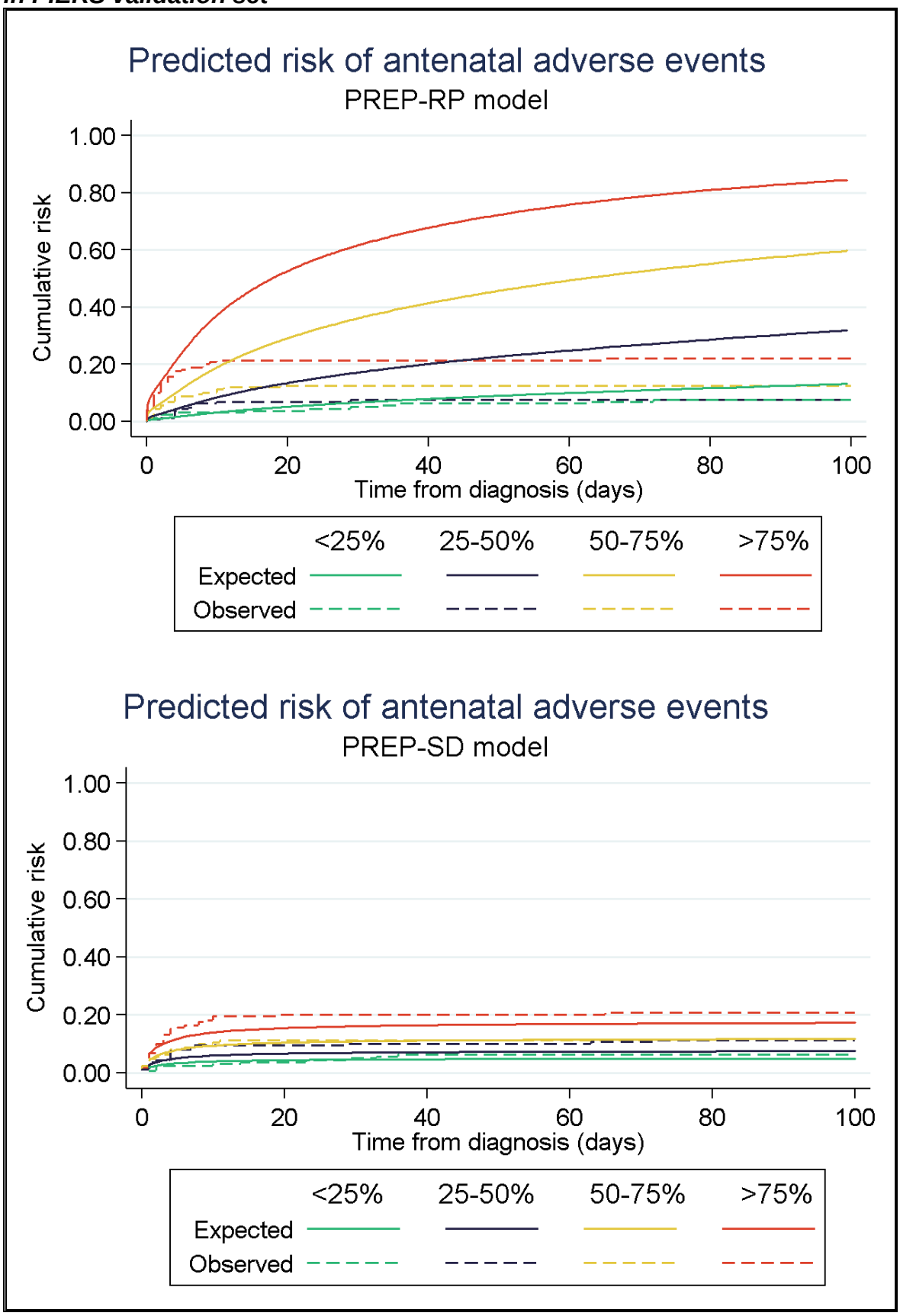
groups at early time points, reflecting the difference in management of competing events between the models. For the two highest risks group (50-75th centile and >75th centile) the risk predictions at four weeks from the *PREP-RP* model are over three times those predicted by the *PREP-SD* model. By this time, the majority of the competing events (delivery without antenatal adverse events) have occurred (see Figure 6.2) resulting in larger levels of competing risks bias.

Table 6.4: Expected proportion of antenatal adverse events by risk groups

	PREP-RP model			**PREP-SD model**		
	2 days	**1 week**	**4weeks**	**2 days**	**1 week**	**4weeks**
<25th centile	1.3%	2.8%	6.7%	2.4%	3.6%	4.4%
25-50th centile	3.5%	7.4%	17.2%	3.6%	5.4%	6.7%
50-75th centile	8.0%	16.6%	35.9%	5.7%	8.4%	10.4%
>75th centile	15.2%	30.1%	58.5%	8.8%	12.8%	15.8%

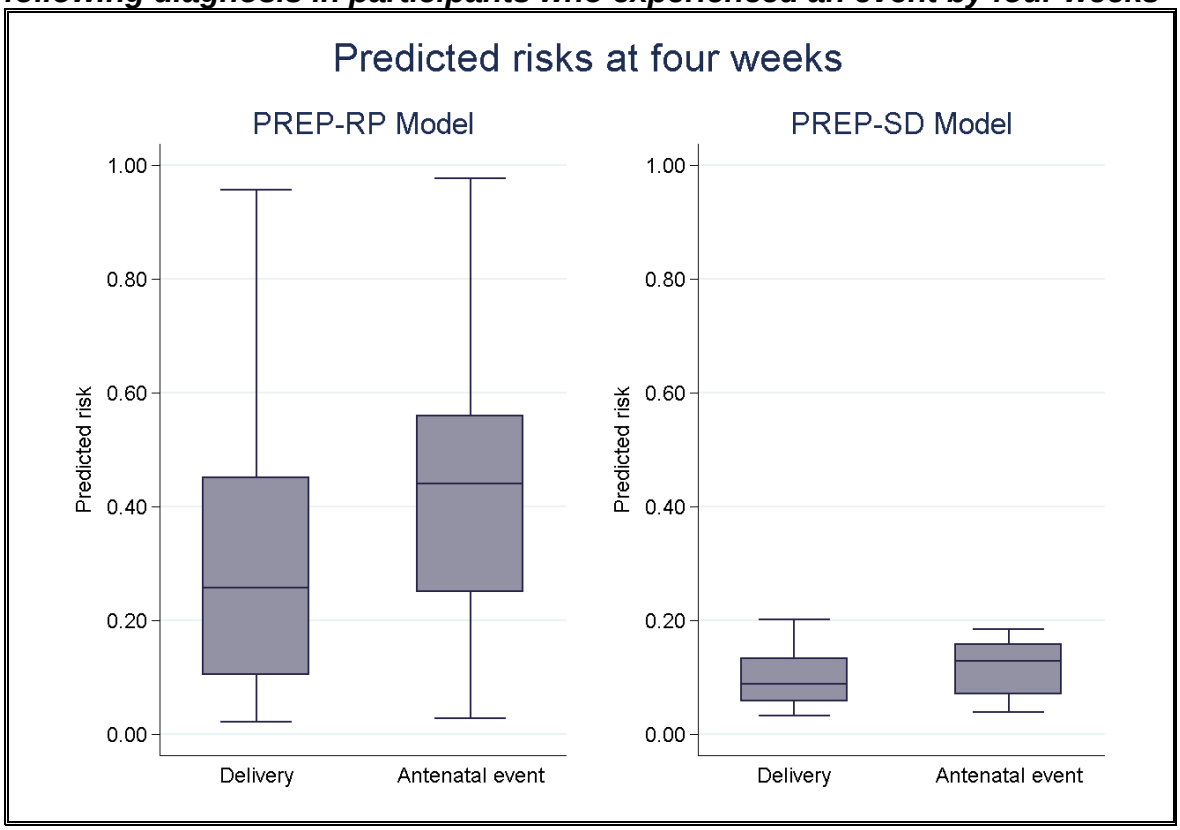
The predicted and observed risks of antenatal adverse events for the risk groups over time are depicted in Figure 6.5. The *PREP-RP* model appears to perform well at early time points, when fewer competing events have occurred. However, at later time points the model greatly overestimates the risks for all four risk groups. At time $t = 100$ the predicted risk (solid line) for the highest risk group (red line) is 84.4% in comparison to an observed risk (dashed lines) of 22.0%, a severe level of competing risks bias. There is little distinction between the observed risk in the lowest two (<25th centile: green line, 25-50th centile: blue line) risk groups. The *PREP-SD* model produces risk predictions that better reflect the observed incidence, though it tends to underestimate the observed risk, particularly in the highest (>75th centile: red line) risk group (this is because the baseline risk is higher in PIERS than in PREP, as shown in Figure 6.2). There is little distinction between the observed risk in the middle (25-50th centile: blue line, 50-75th centile: yellow line) risk groups generated using the *PREP-SD* model.

Figure 6.5: Expected and observed risk of antenatal adverse events by risk group in PIERS validation set



The distribution of the predicted risks at four weeks were compared between participants who had experienced an antenatal adverse event and those who had delivered by four weeks. These distributions are depicted in Figure 6.6. For both prognostic models the distributions overlap, but some separation is observed in median risk estimates. As the absolute risk estimates from the *PREP-SD* model are smaller than those of the *PREP-RP* model, the separation between predictions for the events appears to be less prominent than that observed in the *PREP-RP* predictions.

Figure 6.6: Box and whisker plot of distribution of predicted risks four weeks following diagnosis in participants who experienced an event by four weeks



The discriminative ability of both prognostic models was assessed, the results are provided in Table 4.9. For both models, Harrell's C-index measures are greater at earlier time points. The *PREP-RP* model appears to have an advantage over the *PREP-SD* model in terms of discriminative performance. While overall Harrell's C-index measures were similar, 0.65 for *PREP-RP* and 0.63 for *PREP-SD*, the *PREP-RP* model returns much higher C-index when assessed at two days. The proportion of

explained variation, represented by the R^2_D measures, is 12.4% for the *PREP-RP* model and 11.2% for the *PREP-SD* model, again very similar.

Table 6.5: Measures of discrimination for prognostic models in validation set

		PREP-RP model	PREP-SD model
Harrell's C-index	**2 days**	0.759 (0.67, 0.85)	0.656 (0.57, 0.74)
	1 week	0.704 (0.64, 0.77)	0.658 (0.59, 0.72)
	4weeks	0.683 (0.62, 0.74)	0.649 (0.59, 0.71)
	Overall	0.646 (0.58, 0.71)	0.626 (0.57, 0.69)
D-statistic	**D**	0.771	0.726
	R2D	0.124	0.112

6.4 Discussion

In this chapter two prognostic models, which predict the risk of antenatal adverse events in women diagnosed with early-onset pre-eclampsia, were externally validated in data from an independent study. The *PREP-RP* and *PREP-SD* models were developed earlier in this thesis using different approaches; one using a standard time-to-event approach which ignores the competing event of delivery, the other using a subdistribution approach which appropriately accounts for deliveries. The overall aim of this chapter is to investigate the impact of the management of competing events during prognostic model development on measures of predictive performance. The key findings and conclusions of this external validation study are summarised below.

6.4.1 Key findings

Measures of calibration were used to quantify the accuracy of absolute risk predictions from the prognostic models. The *PREP-RP* model is well calibrated for earlier time-points, this may be may be due to fewer competing events occurring directly after diagnosis (see Figure 6.2). The model is also able to capture the steep increase in risk of antenatal adverse events that occurs directly after pre-eclampsia diagnosis (see Figure 6.5). However, the model overestimates the risk of antenatal adverse events at later time-points, after more competing events have occurred (see Table 6.3). Whereas the *PREP-SD* model underestimates the risk of antenatal adverse events in the validation set at all time points (see Table 6.3); this is more apparent in higher risk groups (see Figure 6.5). The expected risk of antenatal adverse events in the validation set predicted by the *PREP-SD* model better reflects the observed incidence, which account for competing risks, than those predicted by the *PREP-RP* model (see Figure 6.5). The difference in absolute predicted risks from each model is due to differences in the management of competing events during model development. *The PREP-RP* model censors competing events as they occur, and fails to account for

the competing event (delivery) when estimating absolute risks, thus produces inflated absolute risk predictions (competing risks bias). Whereas the *PREP-SD* model appropriately incorporates the competing risks (delivery) when estimating absolute risks, thus produces more representative absolute risk predictions. The underestimation of absolute risks in this study may be due to disparities in the incidence of antenatal adverse events between the development (PREP) and validation (PIERS) studies (see Table 6.1). Model updating and recalibration strategies could be applied to improve both models' predictive performance (Altman et al., 2009, Riley et al., 2016), for example by recalibrating the intercept term in the baseline hazard.

Measures of discrimination were used to quantify how well the models distinguish between those who experience each event, and the time to the event of interest. When assessed in the validation set, the *PREP-RP* model produced higher C-index and D-statistic measures than the *PREP-SD* model (see Table 6.5). It is possible that the wider range of absolute risk predictions from the *PREP-RP* model (see Figure 6.4) give the model an advantage when considering measures of discrimination. Regardless, the discriminative ability of a model does not indicate that a model is suitable for use in clinical practice, particularly when measures of calibration are poor, and the impact of competing risks bias is large, as observed at later time points for the *PREP-RP* model.

6.4.2 Recommendations

This external validation study highlights the impact of the management of competing events during prognostic model development on measures of predictive performance. The *PREP-RP* model, which censors competing events as they occur, performs well at earlier time points where fewer competing events have occurred. The model is able to discriminate between those who experience antenatal adverse events

and those who do not, particularly at the earlier time points (such as 2 days after pre-eclampsia diagnosis). This model may be preferred for assessing the immediate or short-term risks of antenatal adverse events following early-onset pre-eclampsia diagnosis. The *PREP-SD* model, which appropriately accounts for competing events, was expected to produce absolute risk estimates that better reflect the observed risks, particularly at later time points after more competing events had occurred. The model consistently underestimated the absolute risk of antenatal adverse events. However, it is suspected the underestimation is due to disparities in the incidence and recording of the event of interest, rather than attributable to the statistical methods. This model may be preferred for assessing the risks of antenatal adverse events at later time points (such as 1 week after pre-eclampsia diagnosis). However, model updating and re-calibration methods would be advised if using the model in settings other than those in which it was developed.

The external validation exercise also shows the importance of accounting for competing risks during the validation itself; here, the predictive performance measures were derived after accounting for the competing event of delivery. In general, it is recommended that competing risks be appropriately accounted for during both the development and validation of prognostic models by utilising appropriate statistical methods.

However, the impact of competing risks on absolute risk predictions, and thus predictive performance, is associated with the number of observed competing events. If the incidence of observed competing events is small in comparison to the event of interest, the impact on the model's predictive performance will likewise be small. The incidence of delivery in the validation set is lower at earlier time points, when assessed at two days the *PREP-RP* model outperformed the *PREP-SD* model in terms of calibration and discrimination. The amount of competing risks bias in the *PREP-RP* model predictions grows as time progresses and a greater number of competing

events are observed (see Figure 6.5). Thus, it may not always be necessary or impactful to account for competing events during the development and validation of prognostic models.

6.4.3 Limitations and further research

As previously discussed, ideally a minimum of 100 events are required for an external validation study (Collins et al., 2016, Vergouwe et al., 2005). Hence the external validation study in this chapter may be slightly underpowered for unbiased and precise estimation of predictive performance measures, as only 79 events are observed. However, the PIERS cohort was the only one available to validate the models without the need to conduct a lengthy cohort study, which would be considered out of scope for this thesis.

Disparities in the incidence and recording of antenatal adverse events between the development (PREP) and validation (PIERS) studies may have contributed to the difference in predictive performance in this study. A greater proportion of antenatal adverse events were observed in the validation set (see Figure 6.2), however the *PREP-SD* model consistently underestimated the observed risks (see Table 6.3). Abruptions, the most common antenatal adverse event observed in 25 PREP study participants (representing 2.6% of all PREP participants and 33.3% of all observed antenatal adverse events), were not recorded in the PIERS study participants. Thus, participants in the validation set who experienced abruptions may have been misclassified as experiencing a delivery without an adverse event in the external validation analysis. Additionally, as abruptions were the most common antenatal adverse event in the development study, the relationships between the prognostic factors included in both models and the risk of adverse events reflect the risks of these events. Similarly, differences in the incidence of antenatal adverse events in each study are also likely to affect the predictive performance of the models. Despite these

discrepancies, the *PREP-SD* model performs well in the validation set, suggesting the impact of the discrepancies is small. Again, the PIERS cohort was the only one available to validate the models, and a new cohort study, which would take years to conduct, was considered out of scope for this thesis.

In practice, the prognostic models assessed in this external validation study could be utilised alongside clinical expertise to prompt clinical action. For example, in pre-eclampsia high risk patients may be kept in hospital for observation, and delivery of the pre-term baby may be induced in those considered at risk of maternal death (Wilkinson, 2011). Though out of scope for this thesis, further research into the performance of these models could utilise net benefit approaches and decision curves to compare the impact of each model's risk prediction at different time points on clinical actions (Vickers et al., 2015).

Finally, the impact of accounting for competing events in prognostic model development and validation is associated with the incidence of competing events. If the incidence of competing events is low, for example due to short prediction horizon, the impact of not accounting for competing risks is small, and there is little need to account for the competing events. Conversely when the incidence of competing events is high, the impact is large, and it is imperative that competing events are appropriately accounted for. The two models validated in this chapter explore a single competing risk scenario, with a relatively rare event of interest and a relatively common competing event. However, little is known about the impact of the management of competing events on measures of predictive performance under different competing risk scenarios. Current guidance suggests it may only be necessary to account for competing events "when the proportion of subjects experiencing a competing risk is equal or greater to the proportion of subjects experiencing the primary outcome" (Berry et al., 2010). This guidance is often considered as a "rule-of-thumb" for helping researchers decide when competing risks models are needed. The next chapter of this

thesis investigates the impact of accounting for competing events in relation to this rule-of-thumb for a number of competing risks scenarios.

7 WHEN ARE COMPETING RISKS STATISTICAL METHODS NEEDED IN PROGNOSTIC MODEL RESEARCH?

7.1 Introduction

The impact of competing risks statistical methods, during the development of prognostic models for early-onset pre-eclampsia, on measures of predictive performance was examined in the previous chapter. The external validation study identified a substantial difference in predictive performance between the prognostic models developed with and without competing risks statistical methods, when evaluated using a competing risks framework. The difference in predictive performance between the models increased over time, as more competing events occurred. This indicates an association between the impact of the statistical methods and the incidence of both the event of interest and the competing event.

7.1.1 The rule-of-thumb

The scenario studied throughout this thesis (early-onset pre-eclampsia) examines a rare event of interest (antenatal adverse events) and a common competing event (deliveries). This combination is not usually observed in prognostic model research, and the recommendation to always account for competing events, made in earlier chapters, may not be generalizable to all prognostic model scenarios. Indeed, in some circumstances where the incidence of the competing event is low, the effect of not appropriately accounting for competing events may be small and unimportant. It is suggested that it may only be necessary to use the more complex competing risks statistical methods *"when the proportion of subjects experiencing a competing risk is equal or greater to the proportion of subjects experiencing the primary outcome, or*

when follow-up exceeds 5-years." (Berry et al., 2010). This guidance is often considered a "rule-of-thumb" for helping researchers decide when competing risks statistical methods are required, but has received little evaluation.

7.1.2 Simulation studies

Simulation studies are used to assess statistical analysis methods in relation to a known truth (Burton et al., 2006, Morris et al., 2019). Previous simulation studies which compare standard and competing risks time-to-event analyses have focused on assessing differences between cause-specific and subdistribution hazard ratios and regression coefficient estimates (Beyersmann et al., 2009, Dignam et al., 2012, Grambauer et al., 2010, Latouche et al., 2007, Leoce, 2016). However, these measures are not of primary interest in prognostic model research, where the focus is the model's prognostic performance (assessed using measures of calibration and discrimination). For prognostic model research, simulation studies which investigate bias in measures of calibration and discrimination would be more applicable. The calibration of a prognostic model represents the prediction accuracy of the model; how well absolute risk predictions estimated using the model reflect the observed risk in that population on average. Not appropriately accounting for competing events when they are present may lead to inflated absolute risk estimates (competing risks bias), when considering "real-world" risks of an event. The bias is measured as the difference between a non-parametric Kaplan-Meier estimate of the absolute risk of the event of interest and the cumulative incidence at that time. The amount of bias is likely to be affected by the same factors found to cause differences between cause-specific and subdistribution hazard estimates, including those described in the rule-of-thumb (i.e. incidence of the event of interest and competing event). Therefore, this chapter will investigate whether the rule-of-thumb is appropriate for avoiding substantial bias in measures of calibration when developing prognostic models. Investigations into the

effects of competing risks bias on measures of discrimination are not considered in this thesis.

7.1.3 Aims

The overall aim of this chapter is to investigate bias in measures of calibration, specifically bias in cumulative incidence estimates, from prognostic models developed using standard time-to-event methods in the presence of competing events; and hence provide guidance on when competing risks statistical methods are required in prognostic model research.

A simulation study was conducted to assess bias in estimates of the cumulative incidence function across a range of scenarios, in which both the proportion of participants that experience the event of interest and the proportion of participants that experience the competing event are varied. The cumulative incidence function is estimated for a prognostic model developed using standard time-to-event methods. Bias in the cumulative incidence function is calculated as the deviation of these estimates from the “real world” risks, which appropriately account for the competing events. The resulting bias measurements are inspected against the rule-of-thumb to determine whether the rule is appropriate for avoiding substantial bias. Recommendations on the use of competing risks statistical methods for prognostic model research are then provided.

7.2 Methods: Simulating competing risks and evaluating bias

The methods for simulating competing risks data, the simulation scenarios examined in this study, and the simulation procedure are now discussed.

7.2.1 Simulating competing risks data

Methods for generating survival times for time-to-event analyses are summarised by Bender et al. (Bender et al., 2005). These methods were extended by Beyersmann (Beyersmann et al., 2009) to simulate competing risks data using cause-specific hazards and a multi-state framework. Beyersmann ascertains the stochastic behaviour of the competing risks process is completely determined through the cause-specific hazards (Beyersmann et al., 2009). Utilising this, cause-specific hazard functions $h_k(t|\boldsymbol{X})$ are first specified, and then survival times are simulated using the all-cause hazard function $\mathrm{h}(\mathrm{t}|\mathbf{X})$:

$$\mathrm{h}(\mathrm{t}|\mathbf{X}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{h}_{\mathrm{k}}(\mathrm{t}|\mathbf{X}) \qquad \textit{Equation 7.1}$$

A multinomial distribution is then used to decide which of the competing events occurred at each event time T. The probability of an individual experiencing event k given the individual fails at time T is:

$$\mathrm{P}(\mathrm{event} = \mathrm{k}|\mathrm{T} \in \mathrm{dt}, \mathrm{T} \geq \mathrm{t}) = {\mathrm{h}_{\mathrm{k}}(\mathrm{T}|\mathbf{X})}\Big/{\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{h}_{\mathrm{k}}(\mathrm{T}|\mathbf{X})} \qquad \textit{Equation 7.2}$$

In which dt is the infinitesimal interval $[\mathrm{t}, \mathrm{t} + \mathrm{dt})$. The above is equal to the cause-specific hazard at time T divided by the all-cause hazard at time T, or rather the contribution of the cause-specific hazard as a proportion of the total hazard at time T.

The data for this study are simulated using the *survsim* statistical package (Crowther and Lambert, 2013) in Stata 14, which was developed to simulate complex time-to-event data.

7.2.2 Simulation scenarios

All scenarios to investigate the impact of competing events on predictive performance in this simulation study assess the risk of an event of interest in the presence of a single competing event (i.e. $\mathrm{K} = 2$). For simplicity, the cause-specific hazard functions for both the event of interest and the competing event are assumed to remain constant over time (i.e. the time-to-event is exponentially distributed).

It is advised that simulated data should resemble a real study in order for the results to be generalizable and have credibility (Burton et al., 2006). A recent study which developed a competing risks prognostic model for coronary risk prediction (Wolbers et al., 2009) is used as a motivating example, and provides parameter values to define the "*base scenario*" in this simulation study. Wolbers et al. predict the ten-year risk of coronary heart disease (CHD) in 4,144 women, 465 (11.2%) of which experienced a CHD event (event of interest) and 1,263 (30.5%) experienced a non-CHD death (competing event) during follow-up (10 years).

To assess the rule of thumb proposed by Berry (Berry et al., 2010), the proportion of participants who experience each event is required (i.e. cause-specific cumulative incidences, $F_k(t)$). The rule-of-thumb advises the use of competing risk statistical methods to account for competing events when the proportion of competing events is at least that of the event of interest. Thus, a constant scaling factor $\gamma > 0$ is introduced to determine the relationship between the proportion of participants who experience the event of interest:

$$\gamma \mathrm{F}_1(\mathrm{t}) = \mathrm{F}_2(\mathrm{t}), \;\; 0 < \{\mathrm{F}_1(\mathrm{t}) + \mathrm{F}_2(\mathrm{t})\} \leq 1 \qquad \textbf{\textit{Equation 7.3}}$$

Competing risk statistical methods should be used when $\gamma \geq 1$.

For the base scenario of this simulation study, the cumulative incidence for the event of interest $\mathrm{F}_1(10)$, is 0.112 and the cumulative incidence for the competing event $\mathrm{F}_2(10)$, is 0.335 resulting in a scaling factor $\gamma = {}^{0.335}/_{0.112} = 2.72$. The proportion

of participants who experience competing events is 2.72 times the proportion who experience the event of interest. In this instance, the rule-of-thumb would advise the use of competing risks statistical methods to account for the competing events.

Within this simulation study, a range of plausible scenarios are investigated with varying event of interest and competing event incidences. The event of interest incidences were selected to reflect a range from rare events $(\mathrm{F}_1(10) = 0.05)$ to more common events $(\mathrm{F}_1(10) = 0.50)$, with values of γ ranging from 0.25 to 2.72 (observed by (Wolbers et al., 2009). The scenarios are specified in detail in Table 7.2.

Previous simulation studies which compare standard and competing risks time-to-event analyses (Beyersmann et al., 2009, Dignam et al., 2012, Grambauer et al., 2010, Latouche et al., 2007, Leoce, 2016) found that uninformative censoring did not introduce bias or significantly alter study findings. Thus, censoring was not introduced during simulation study follow-up. However, a maximum follow-up time of $\mathrm{t} = 10$ was enforced and any individual who had not experienced either event was censored at that time.

7.2.3 The simulation process

In order to investigate bias in measures of calibration for each scenario listed above, the following simulation process outlined in Table 7.1 was applied.

Table 7.1: The simulation process for evaluating overall calibration bias

Step 1	Select one of the scenarios from Table 7.2.
Step 2	Determine the cause-specific hazards for the event of interest and the competing event, using Equation 7.4.
Step 3	Simulate a single study containing 4,144 participants using the survsim command in Stata.
Step 4	Calculate the naïve cumulative incidence and difference in cumulative risk.
Step 5	Repeat steps 3 & 4 x500 times.
Step 6	Evaluate cumulative incicdence bias.

The simulation process is discussed in more detail below.

7.2.3.1 Step 1) Define the scenario

Define the proportion of participants that experience the event of interest by time 10, $F_1(10)$, and the scaling factor for the relative proportion of competing events, γ (i.e. select one of the scenarios from Table 7.2).

7.2.3.2 Step 2) Determine cause-specific hazards

In order to simulate competing risks time-to-event data using an exponential model, the constant cause-specific hazards for both the event of interest and competing event are required. Given the proportion of participants who experience the event of interest $F_1(t)$, and the scaling factor γ, and utilising exponential models with constant hazards; $h_1(t) = \alpha_1$ for the event of interest, and $h_2(t) = \alpha_2$ for the competing event; it is possible to determine the constant cause-specific hazard term for both the event of interest and the competing event:

$$F_1(t) = \int_0^t \alpha_1 \exp\{-(\alpha_1 + \alpha_2)s\}\, ds$$

$$= \alpha_1 \left[\frac{\exp\{-(\alpha_1 + \alpha_2)t\}}{-(\alpha_1 + \alpha_2)}\right]_0^t$$

$$= \alpha_1 \left[\frac{\exp\{-(\alpha_1 + \alpha_2)t\}}{-(\alpha_1 + \alpha_2)} - \frac{1}{-(\alpha_1 + \alpha_2)}\right]$$

$$F_1(t) = \frac{\alpha_1}{\alpha_1 + \alpha_2}[1 - \exp\{-(\alpha_1 + \alpha_2)t\}] \qquad \textit{Equation 7.4}$$

The derivation of this equation is provided in Appendix X.

For example, in the base scenario the cumulative incidence for the event of interest $F_1(10) = 11.2\%$ and the scaling factor $\gamma = 2.72$, the corresponding cause-specific hazards are calculated as follows:

$\alpha_1 = -\frac{\ln\left[1 - \frac{1+2.72}{2.72} \times 0.305\right]}{(1+2.72)10} = 0.01451$, and $\alpha_2 = 2.72 \times 0.01451 = 0.03947$

7.2.3.3 Step 3) Generate time-to-event data for a single study

Generate a single study containing 4,144 participants (the sample size in (Wolbers et al., 2009)), with time-to-event and event indicator variables simulated with exponential distributions for both the event of interest and competing event. Specify the cause-specific hazards as calculated in Step 2 and restrict follow-up time to a maximum of $t = 10$. Study data are simulated using the *survsim* command in Stata.

7.2.3.4 Step 4) Calculate naïve cumulative incidence and difference in cumulative risk

Using the simulated study data, calculate the non-parametric Kaplan-Meier estimate of the cumulative incidence, naïve to the occurrence of competing events, at $t = 10$. Determine the difference in cumulative risk as the difference between the Kaplan-Meier estimate at $t = 10$ and the specified cumulative incidence $F_1(10)$, based on the true model.

7.2.3.5 Step 5) Repeat simulations

Repeat steps three and four 500 times to obtain the distribution of bias in estimates of the cumulative risk for the specified scenario.

7.2.3.6 Step 6) Evaluate cumulative incidence bias

Summarise the distribution of bias in estimates of the cumulative risk across the 500 simulations by calculating summary statistics for absolute measures (mean, SD) and measures relative to the expected cumulative incidence $F_1(10)$ as specified in Step 1 (mean percentage).

An example of the Stata code used to simulate the data and evaluate cumulative incidence bias is provided in Appendix XIII.

The results for each scenario will be considered alongside the rule-of-thumb to scrutinise whether the rule is appropriate for avoiding substantial cumulative incidence

bias in prognostic model studies. Differences in estimates of the cumulative incidence function are shown graphically by plotting Kaplan-Meier estimates of cumulative risk, naïve to the occurrence of competing events, and the non-parametric cumulative incidence estimates, which account for the competing events, for the first 30 simulated studies for each scenario.

7.3 Results: Simulating competing risks and evaluating bias.

The results of the simulation study to investigate bias in measures of calibration, from prognostic models developed using standard time-to-event methods in the presence of competing events, are now discussed.

7.3.1 Determining cause-specific hazards

For each scenario, the constant cause-specific hazard for both the event of interest and the competing event were calculated using Equation 7.4. The resulting hazards are reported in Table 7.2.

7.3.2 Evaluating overall calibration bias

For the first 30 simulated studies in each scenario, Kaplan-Meier estimates of the cumulative incidence of the event of interest, naïve to the occurrence of competing events, are depicted alongside non-parametric cumulative incidence estimates, which account for competing events. The resulting graphs are depicted in Figure 7.1, Figure 7.2, Figure 7.3, and Figure 7.4. The difference between the Kaplan-Meier (blue lines) and cumulative incidence (red lines) estimates reflects the impact of not accounting for competing events when estimating absolute risks for the event of interest.

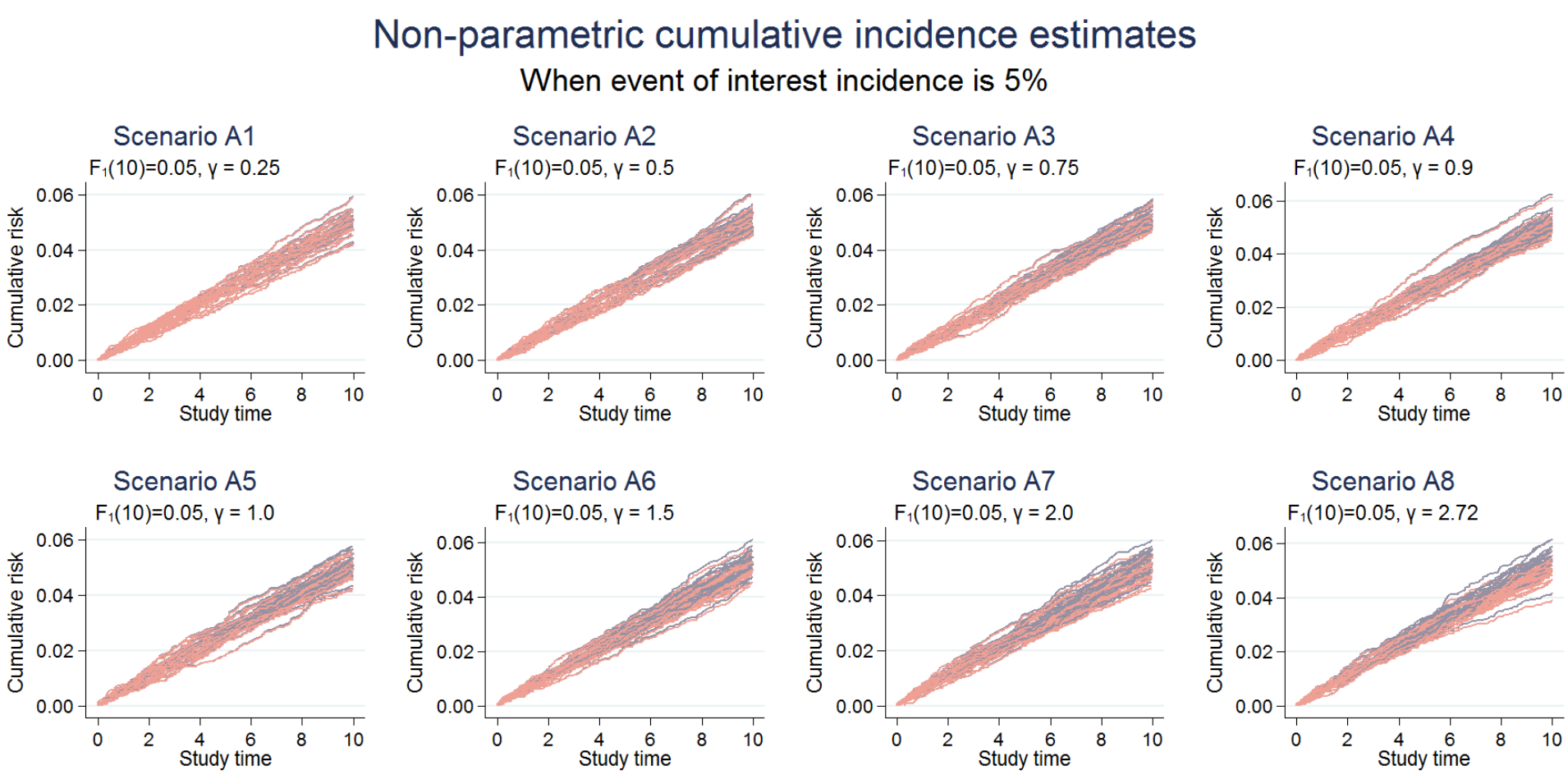
Separation between the non-parametric estimates becomes more apparent as the event of interest incidence $F_1(t)$ and relative proportion of competing events γ increase, i.e. a greater number of competing events are observed. In scenarios where there is little observable separation between the non-parametric estimates (i.e. scenarios A1-A8, B1-B6, and C1-C4), applying standard time-to-event analysis methods will give similar cumulative risk estimates to an analysis using competing risks analysis methods. Whereas, substantial differences are likely to be observed in cumulative risk estimates resulting from those scenarios with notable separation (i.e. scenarios B7-B8, C5-C6, and D1-D4).

Table 7.2: Constant cause-specific hazards for event of interest and competing event for scenarios investigated in this simulation study

Scenario	Incidences		Scaling factor γ	Cause-specific hazards	
	Event of interest $F_1(10)$	Competing event $F_2(10)$		Event of interest α_1	Competing event α_2
A1	5.0%	1.25%	0.25	0.00516	0.00129
A2		2.5%	0.50	0.00520	0.00260
A3		3.75%	0.75	0.00523	0.00392
A4		4.5%	0.90	0.00525	0.00473
A5		5.0%	1.00	0.00527	0.00527
A6		7.5%	1.50	0.00534	0.00801
A7		10.0%	2.00	0.00542	0.01083
A8		13.6%	2.72	0.00553	0.01505
B1	11.2%	2.8%	0.25	0.01207	0.00302
B2		5.6%	0.50	0.01226	0.00613
B3		8.4%	0.75	0.01247	0.00935
B4		12.4%	0.90	0.01259	0.01133
B5		11.2%	1.00	0.01268	0.01268
B6		16.8%	1.50	0.01314	0.01971
B7		22.4%	2.00	0.01365	0.02730
B8: Base Scenario		30.5%	2.72	0.01449	0.03946
C1	20.0%	5.0%	0.25	0.02301	0.00575
C2		10.0%	0.50	0.02378	0.01189
C3		15.0%	0.75	0.02462	0.01846
C4		18.0%	0.90	0.02516	0.02264
C5		20.0%	1.0	0.02554	0.02554
C6		30.0%	1.5	0.02773	0.04159
C7		40.0%	2.0	0.03054	0.06109
C8		54.4%	2.72	0.03663	0.09963
D1	50.0%*	12.5%	0.25	0.07847	0.01962
D2		25.0%	0.50	0.09242	0.04621
D3		37.5%	0.75	0.11883	0.08912
D4		45.0%	0.90	0.15767	0.14190

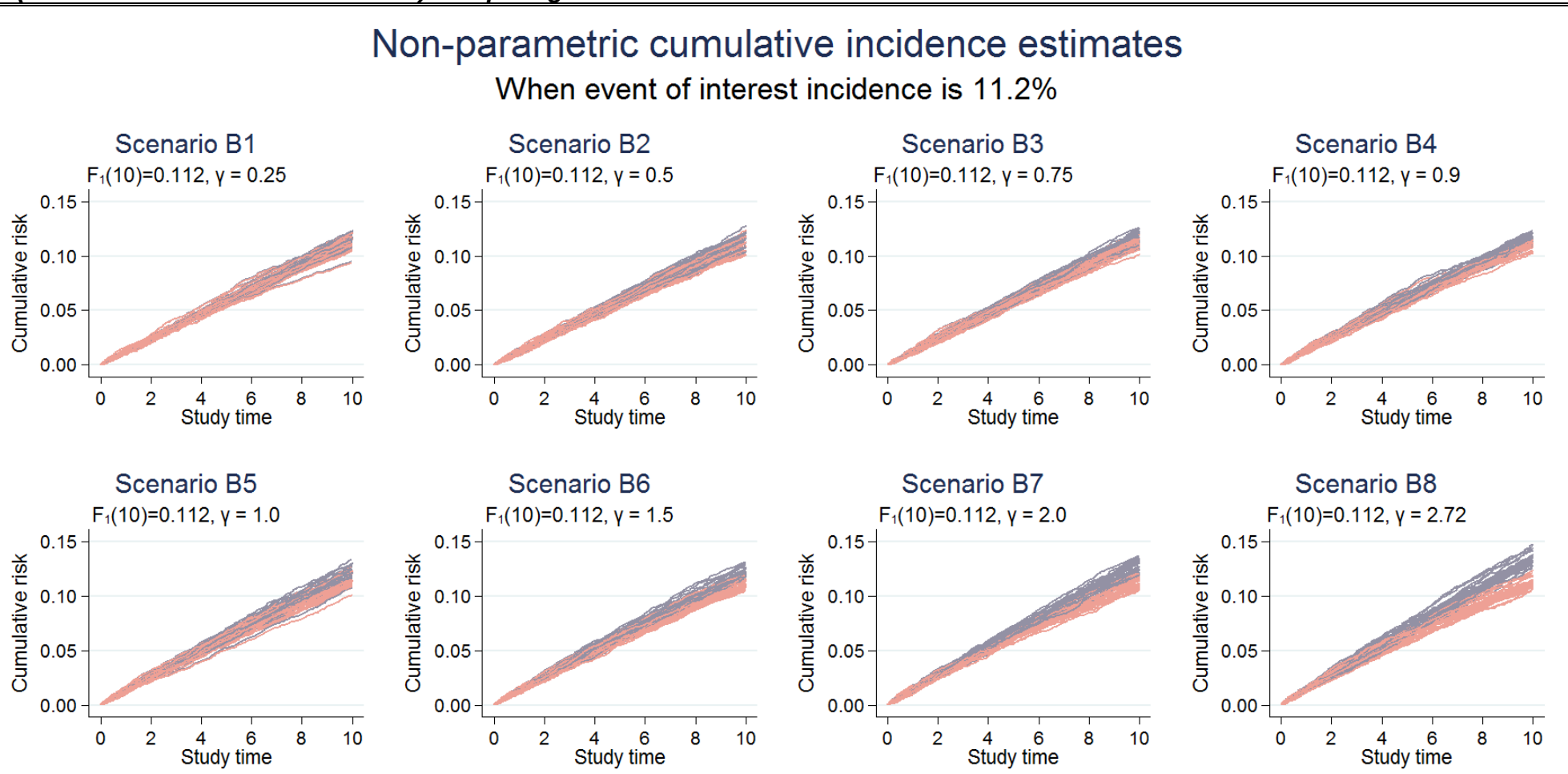
***It is impossible for the sum of the proportion of events of interest $(F_1(10))$ and competing events $(F_2(10))$ to exceed 100%, thus Scenario D $(F_1(10) = 50\%)$ is limited to instances where $\gamma \leq 1$.**

Figure 7.1: Difference in non-parametric estimates of cumulative incidence when ignoring (Kaplan-Meier – blue lines) and accounting for (cumulative incidence – red lines) competing events in simulation scenarios with 5% event of interest incidence.



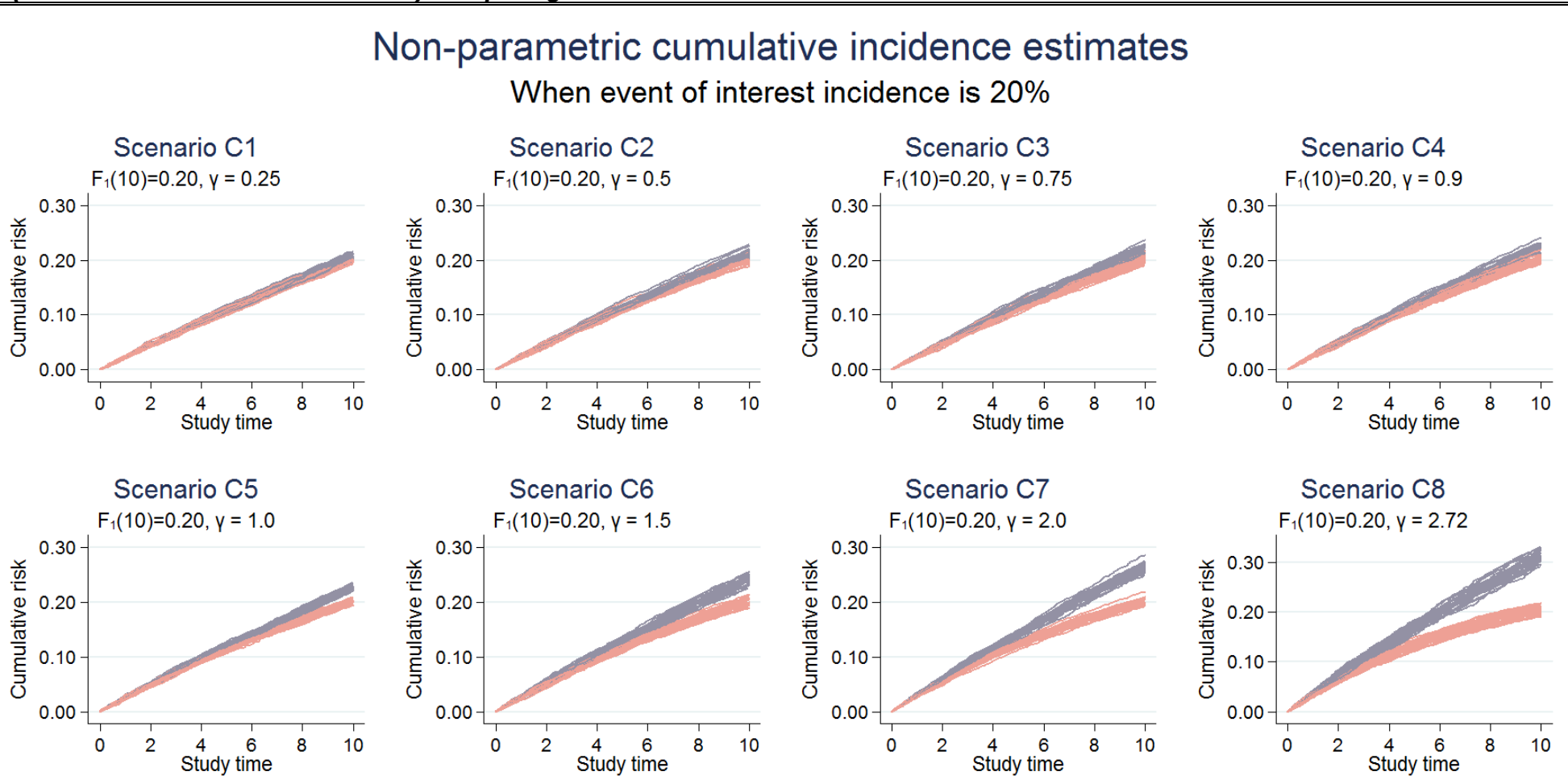
Blue lines represent Kaplan-Meier estimates which censor the competing events.
Red lines represent cumulative incidnce function estimates which appropriately account for the competing events.

Figure 7.2: Difference in non-parametric estimates of cumulative incidence when ignoring (Kaplan-Meier – blue lines) and accounting for (cumulative incidence – red lines) competing events in simulation scenarios with 10% event of interest incidence.



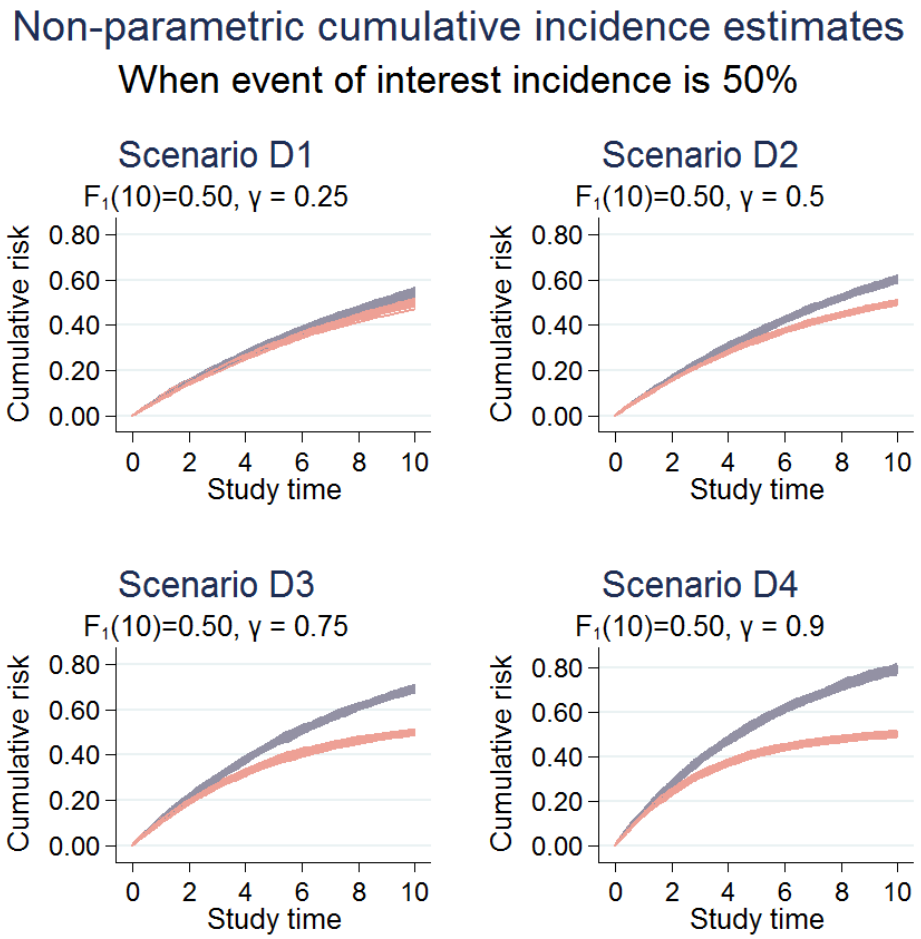
Blue lines represent Kaplan-Meier estimates which censor the competing events.
Red lines represent cumulative incidnce function estimates which appropriately account for the competing events.

Figure 7.3: Difference in non-parametric estimates of cumulative incidence when ignoring (Kaplan-Meier – blue lines) and accounting for (cumulative incidence – red lines) competing events in simulation scenarios with 20% event of interest incidence.



Blue lines represent Kaplan-Meier estimates which censor the competing events.
Red lines represent cumulative incidnce function estimates which appropriately account for the competing events.

Figure 7.4: Difference in non-parametric estimates of cumulative incidence when ignoring (Kaplan-Meier – blue lines) and accounting for (cumulative incidence – red lines) competing events in simulation scenarios with 50% event of interest incidence.



Blue lines represent Kaplan-Meier estimates which censor the competing events.
Red lines represent cumulative incidnce function estimates which appropriately account for the competing events.

Summary measures for cumulative risk bias in each simulation scenario are provided in Table 7.3. It is again evident that both measures of absolute and percentage bias escalate as the event of interest incidence $F_1(t)$ and relative proportion of competing events γ increase.

7.3.3 Evaluation of the rule-of-thumb

The rule-of-thumb (Berry et al., 2010) advises competing risk analysis methods need only be applied when the incidence of the competing event is at least as large as the event of interest, i.e. when $\gamma \geq 1$ (scenarios A5-A8, B5-B8, and C5-C8). Thus, advises the use of standard time-to-event analysis methods for scenarios where the incidence of the competing event is less than the event of interest (i.e. scenarios A1-A4, B1-B4, C1-C4, and D1-D4).

The above advice seems to be appropriate for avoiding substantial bias in a number of the simulation scenarios investigated. Nonetheless, the advice seems less appropriate for scenarios with relatively rare (5%) or common (50%) incidences for the event of interest. In scenarios A5-A6, the rule would advise the use of the more complex methods to account for the competing events. However, the bias in cumulative incidence estimates resulting from the application of standard time-to-event methods in these scenarios is small (ranging from 0.001 to 0.004) and potentially unimportant. Conversely, in scenarios D1-D4 the rule suggests competing risk analysis methods need not be applied to account for the competing events. However, the bias in cumulative incidence estimates in these scenarios is large (ranging from 0.043 to 0.294) and likely to be important.

Table 7.3: Summary of cumulative risk bias in simulation study scenarios

Scenario	Incidence of event of interest $F_1(10)$	Scaling factor γ	Kaplan-Meier estimate $\widehat{F_1}(10)$ Mean (SD)	Absolute bias $\widehat{F_1} - F_1$ Mean (SD)	Percentage bias $(\widehat{F_1} - F_1)/F_1$ Mean
A1	5.0%	0.25	0.051 (0.003)	0.001 (0.003)	1.02%
A2		0.50	0.051 (0.003)	0.001 (0.003)	1.37%
A3		0.75	0.051 (0.003)	0.001 (0.003)	2.17%
A4		0.90	0.051 (0.003)	0.001 (0.003)	2.40%
A5		1.00	0.051 (0.003)	0.001 (0.003)	2.47%
A6		1.50	0.052 (0.004)	0.002 (0.004)	3.80%
A7		2.00	0.053 (0.004)	0.003 (0.004)	5.71%
A8		2.72	0.054 (0.004)	0.004 (0.004)	7.51%
B1	11.2%	0.25	0.113 (0.005)	0.001 (0.005)	1.21%
B2		0.50	0.116 (0.005)	0.004 (0.005)	3.19%
B3		0.75	0.117 (0.005)	0.005 (0.005)	4.09%
B4		0.90	0.118 (0.005)	0.006 (0.005)	5.36%
B5		1.00	0.119 (0.005)	0.007 (0.005)	6.51%
B6		1.50	0.124 (0.006)	0.012 (0.006)	10.48%
B7		2.00	0.128 (0.006)	0.016 (0.006)	13.98%
B8: Base Scenario		2.72	0.135 (0.006)	0.023 (0.006)	20.32%
C1	20.0%	0.25	0.206 (0.006)	0.006 (0.006)	2.81%
C2		0.50	0.212 (0.007)	0.012 (0.007)	5.91%
C3		0.75	0.219 (0.007)	0.019 (0.007)	9.47%
C4		0.90	0.223 (0.007)	0.023 (0.007)	11.48%
C5		1.0	0.226 (0.007)	0.026 (0.007)	12.81%
C6		1.5	0.242 (0.007)	0.042 (0.007)	21.00%
C7		2.0	0.263 (0.008)	0.063 (0.008)	31.63%
C8		2.72	0.307 (0.009)	0.107 (0.009)	53.57%
D1	50.0%*	0.25	0.543 (0.008)	0.043 (0.008)	8.69%
D2		0.50	0.603 (0.008)	0.103 (0.008)	20.63%
D3		0.75	0.695 (0.01)	0.195 (0.01)	39.03%
D4		0.90	0.794 (0.01)	0.294 (0.01)	58.77%

***It is impossible for the sum of the proportion of events of interest $(F_1(10))$ and competing events $(F_2(10))$ to exceed 100%, thus Scenario D $(F_1(10) = 50\%)$ is limited to instances where $\gamma \leq 1$.**

7.4 Extension to assess overall calibration bias over all possible values of γ

The results from the above simulation study provoked further investigation into the levels of cumulative risk bias resulting from applying standard time-to-event analysis methods in the presence of competing risks.

7.4.1 The simulation process

This second simulation study process followed many of the same steps as the one previously, with the following alterations:

7.4.1.1 Step 1) New scenarios

Scenarios with event of interest incidences, $F_1(10)$, equal to 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% were investigated. For each event of interest incidence, cumulative risk bias was assessed for all possible values of γ, divided into 200 equally spaced intervals[i].

7.4.1.2 Step 3) Study size

The number of participants simulated in each scenario was increased to 1,000,000 to reduce the variance in bias estimates due to sample size.

7.4.1.3 Step 5) No repetitions

Following investigations of the increased study sample size in Step 3, the study team felt it unnecessary to repeatedly simulate the scenarios due to the small amount of variance in the simulation estimates.

7.4.2 Results

[i] As $\gamma F_1(t) = F_2(t)$, $0 < \{F_1(t) + F_2(t)\} \leq 1$, (Equation 7.4) it follows $0 < \gamma \leq \frac{1}{F_1(10)} - 1$.

Absolute bias in cumulative incidence estimates for each scenario, over all possible values of γ are displayed in Figure 7.5. Again, the amount of cumulative risk bias increased as the incidence of the event of interest and the relative proportion of competing events increased. This is due to the increase in the incidence of the competing event (a graph depicting bias related to competing event incidence is reported in Appendix XIV). The amount of absolute bias is restricted; in a scenario with an event of interest incidence of 80%, the absolute bias cannot be more than 100%-80%=20%. To observe 0.1 bias in the cumulative incidence estimate when the event of interest incidence is 90%, the competing event incidence need only be $\gamma F_1(10) = 0.11 \times 0.9 = 9.9\%$. Whereas, to observe the same bias with an event of interest incidence of 5%, the competing event incidence needs to be $\gamma F_1(10) = 18.2 \times 0.05 = 91\%$. Still, an absolute bias of 0.1 is more substantial when the expected event of interest incidence is 0.05 compared to when it is 0.90. In prognostic model research, this level of bias would result in predicted absolute risks for the event of interest estimated as 0.15, a threefold increase on observed risks. In prognostic model research, measures of percentage bias may be considered more intuitive. Bias in overall calibration, relative to the event of interest incidence, for all possible values of γ are displayed in Figure 7.6 (a graph depicting bias related to competing event incidence is reported in Appendix XIV).

Figure 7.5: Absolute bias in cumulative incidence estimates for all possible values of γ with $F_1(10)$ from 5% to 90%

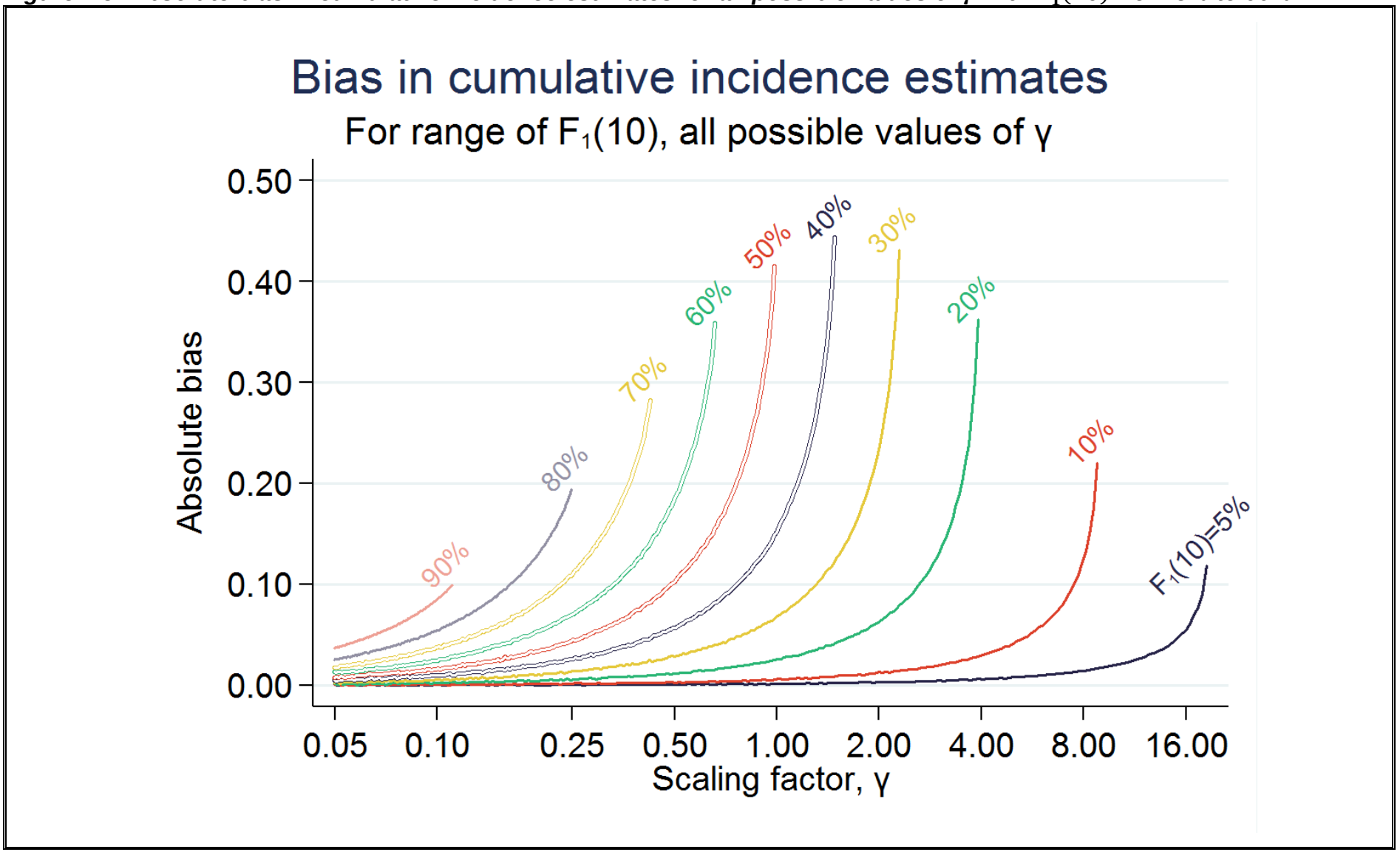
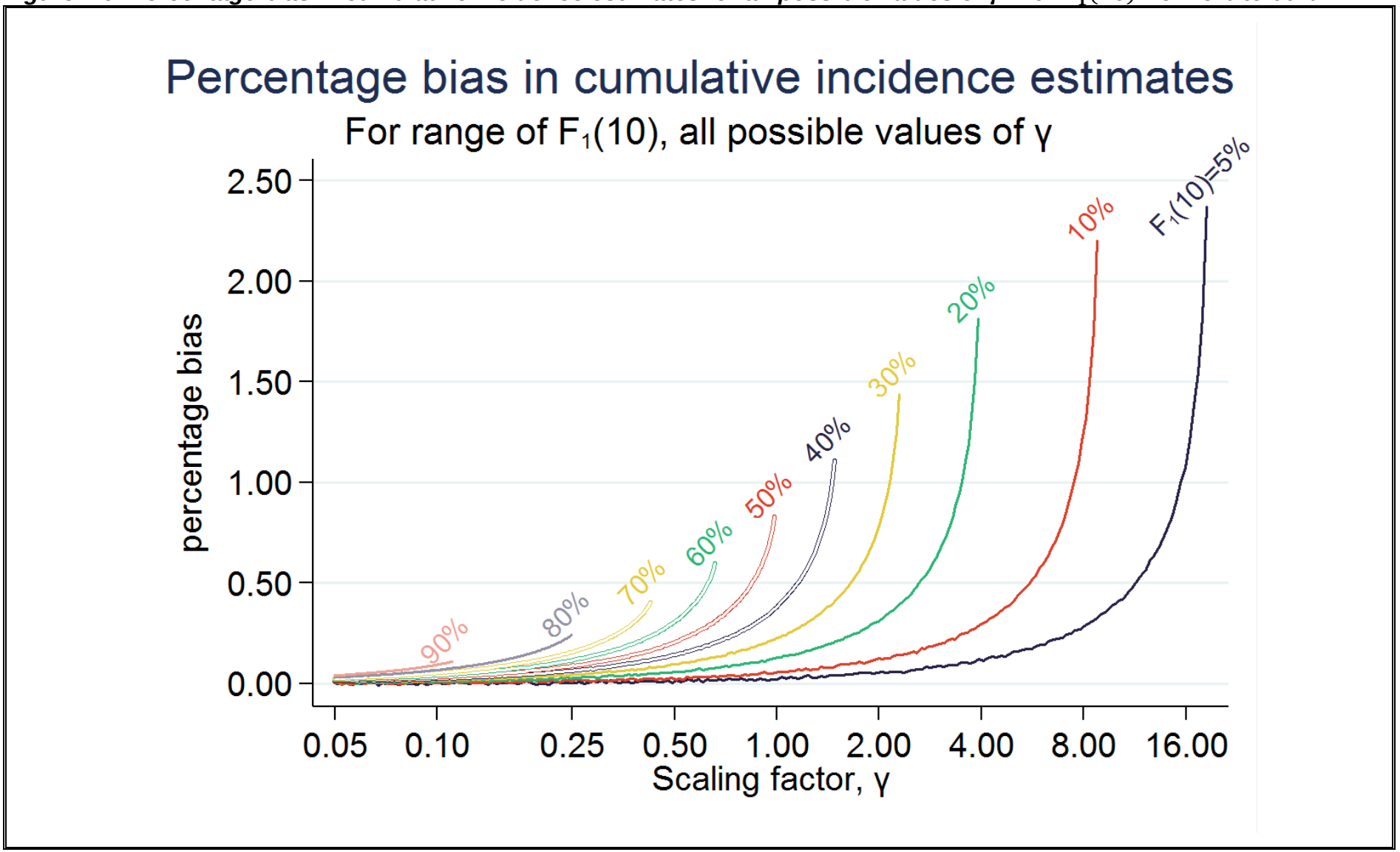


Figure 7.6: Percentage bias in cumulative incidence estimates for all possible values of γ with $F_1(10)$ from 5% to 90%



7.4.3 Evaluation of the rule-of-thumb

The rule-of-thumb (Berry et al., 2010) considers the incidence of the event of interest and the competing event, and suggests actions based on $\gamma \geq 1$. This would advise the use of standard time-to-event analysis methods for scenarios on the left-hand side of Figure 7.5 and Figure 7.6, and advise competing risks analysis methods for scenarios on the right-hand side of the graphs. It is evident that this rule could result in substantial cumulative risk bias if applied in some circumstances; Such as when event of interest incidence is high (e.g. 60%) and the proportion of participants not expected to experience either event is low (e.g. competing event incidence is 30%, resulting in only 10% not experiencing either event).

7.5 Discussion

The simulation studies performed in this chapter assess the bias in measures of overall calibration when standard time-to-event analysis methods are applied to scenarios with competing events. A number of factors, including the incidence of the event of interest and the incidence of the competing event, were varied in scenarios within the simulations to investigate associations with bias in cumulative incidence estimates. A rule-of-thumb was examined to determine whether it is appropriate to apply in prognostic model research to avoid substantial calibration bias. The key findings and conclusions of these simulation studies are summarised below.

7.5.1 Key findings and recommendations

Bias in cumulative incidence estimates was found to be strongly associated with the incidence of both the event of interest and the competing event. The two simulation studies conducted in this chapter found systematic differences in calibration of risk predictions when estimated using standard time-to-event analysis methods rather than competing risk modelling. The bias in cumulative incicdence estimates, which overestimates the absolute risk of the event of interest, increases as the event of interest incidence and the relative proportion of competing events increase. Statistical methods which appropriately account for the competing events can be applied to avoid this bias. However the added complexity of the analyses, interpretation, and communication of the hazard and cumulative incidence functions (which have prompted a number of tutorials including: (Allignol et al., 2011, Andersen et al., 2012, Dignam et al., 2012, Wolbers et al., 2009)) may be off-putting for researchers.

Current guidance on when the more complex competing risks methods should be applied (Berry et al., 2010), was found not to be appropriate for avoiding substantial cumulative risk bias. The rule advised against the use of competing risks methods when the competing event was not as prevalent as the event of interest. This advice

results in high levels of bias in cumulative incidence estimates in scenarios with large event of interest incidences, such as in scenarios D1-D4. The rule also suggests using competing risks methods when the competing event is at least as prevalent as the event of interest. This advice made negligible differences to estimates of calibration in scenarios with a small event of interest incidence, such as in scenarios A5-A8. The existing rule was thus found to be incompatible with avoiding biases in estimates of cumulative risk. The primary purpose of prognostic model research is the accurate prediction of the absolute risks of the event of interest, thus large biases in measures of calibration are unacceptable.

Though bias was observed in all of the scenarios investigated in the initial simulation study (Table 7.3), some researchers may determine the amount of bias to be clinically negligible and thus choose to apply standard time-to-event analysis methods. For example, when the observed event of interest incidence is 11.2% (scenarios B1-B8), a cumulative risk estimate equal to 11.9% (a 6.5% percentage increase due to bias, scenario B5) may be sufficiently adequate, to avoid using the more complex competing risks methods. However, a cumulative risk estimate equal to 13.5% (a 20.3% increase due to bias, scenario 8) may be unsatisfactory, thus the more complex analyses are justified.

A suggested update to the existing rule could utilise these thresholds for acceptable levels of percentage bias. Competing risks methods should be applied when the expected value of percentage bias exceeds a determined threshold ε, and need not be applied if the expected bias is less than the threshold. Table 7.4 displays the maximum competing event incidence to prompt the use of competing risks methods for event of interest incidences of 5, 20, 50, and 80%, if a threshold of $\varepsilon = 0.10$ is accepted. This allows the cumulative incidence estimate to be up to 1.10 times the expected value.

Table 7.4: Limits for updated rule for competing risks methods with ε=0.1

Incidences		Absolute bias	Scaling factor
Event of interest $F_1(10)$	**Event of interest** $F_2(10)$	$\varepsilon \times F_1(10)$	γ
5%	17%	0.5%	3.40
20%	16%	2.0%	0.80
50%	14%	5.0%	0.28
80%	11%	8.0%	0.135

For example, when the incidence of the event of interest is 5% the recommendation would be to apply competing risks methods if the bias in absolute risk predictions exceeds 0.5% (10% of 5%). This occurs when the expected competing event incidence is at least 17% (3.4 times greater). However, when the incidence of the event of interest is 50% the new rule advises applying competing risks methodology once bias in absolute risk predictions exceeds 5%, which occurs when the competing event incidence exceeds 14% (0.28 times greater). The guidance takes into account differences in absolute risk predictions from standard and competing risks statistical approaches, and can be used as a guide by researchers wanting to develop accurate prognostic models in the presence of competing events.

7.5.2 Limitations and further research

All simulation studies within this chapter were developed with constant cause-specific hazard functions through use of the exponential time-to-event model. This assumption was considered sufficient for these initial investigations into the effects of competing events on measures of prognostic model calibration. Furthermore, solving the integral to obtain the cumulative hazard function is straightforward when the hazard function is constant; even after combining multiple cause-specific hazards to obtain the all-cause hazard (Equation 7.4). Additionally, inverting the cumulative hazard function to allow simulation of survival times is also straightforward when an exponential model is used. However, the assumption of constant hazards is often considered not to be biologically plausible in medical research, where often at least one turning point in the hazard is observed over time (Crowther and Lambert, 2013). Their article discusses

methods to obtain the cumulative hazard function when the integral is analytically intractable (numerical integration techniques). As well as methods to solve the equation to generate survival times when the cumulative hazard function is not invertible (iterative root finding methods). Thus, extensions to this research include incorporating more complex hazard functions, through flexible parametric modelling of the underlying hazard functions, to increase the biological plausibility of the models and check the recommendations are consistent with the current findings. Other potential extensions to this work include the incorporation of time-dependent prognostic factor associations and prognostic factors which vary over time.

The included simulation studies utilise non-parametric estimates of the cumulative risk of the event of interest. However, prognostic models apply regression modelling techniques to incorporate associations between multiple prognostic factors and the event of interest (Steyerberg et al., 2013). A recent literature review of competing risks simulation studies (Leoce, 2016) identified studies which investigated varying strengths of associations between a single binary exposure variable and both the event of interest and a single competing event using a range of scenarios (Dignam et al., 2012, Grambauer et al., 2010, Latouche et al., 2007). The studies found the difference between the regression coefficient estimates to be negligible when the exposure is only associated with the event of interest and not the competing event. However, when the exposure is associated with an increase in the risk of event of interest but a reduction in the risk of the competing event, attenuated standard time-to-event analysis regression coefficient estimates were observed. The focus of these previously published studies was to compare cause-specific and subdistribution regression coefficient estimates. However, little is known regarding how these differences, caused by varying associations between a prognostic factor and the competing event, affect prognostic model calibration. Thus, further research is needed

to investigate the effect of varying associations between prognostic factors and the competing event on prognostic model calibration.

8 DISCUSSION

This chapter concludes the thesis with a general discussion of the research presented in previous chapters. Limitations of the research are presented alongside suggestions of future work in the area.

8.1 Thesis Overview

Prognostic model research is important as it aids understanding of the risks of future health outcomes in individuals. Communicating these risks enables clinicians to help their patients understand the likely course of their disease or condition, enhancing the process of informed decision making (Steyerberg et al., 2013). The time to a given health outcome can be of interest in prognostic model research, and thus time-to-event regression methods are often utilised to develop prognostic models. It is often the case that competing events prevent or alter the risk of the event of interest from occurring. If not appropriately accounted for, the presence of these competing events can result in inflated absolute risk estimates, affecting the predictive accuracy of a prognostic model. Though the statistical methods which account for competing risks theory has existed since 1760 (Bernoulli, 1760), there is increasing evidence that the methods are not being utilised (Austin and Fine, 2017, Koller et al., 2012, Schumacher et al., 2016, Walraven and McAlister, 2016), and there is relatively little empirical research which investigates the presence and impact of competing events in prognostic model research.

The overall aim of previous chapters was to improve understanding of the influence of competing risks issues in prognostic model research. In particular, early chapters (Chapters 2 and 3) investigated the presence, reporting, and management of competing events in existing prognostic model research, focusing on systematic

reviews and development articles. In the middle section of this thesis (Chapters 4 to 6), flexible parametric prognostic models for the risks of antenatal adverse events in pre-eclampsia patients were developed and externally validated using standard and competing risks statistical methods. In the penultimate chapter (Chapter 7), the impact of competing risks on prognostic performance measures were evaluated in a competing risks simulation study. A short summary of the chapters contained in this thesis is provided below:

8.1.1 Chapter 2: An empirical evaluation of the presence and reporting of competing events in systematic reviews of prediction model studies.

The aim of this chapter was to empirically investigate the presence and reporting of competing events in published systematic reviews of prediction model development studies. The empirical review identified 31 systematic review articles, published between January 2015 and April 2017, which were assessed on the potential for competing risks bias and acknowledgement of competing events. The key findings was that the majority (61.3%) of the systematic review articles in the review were classified as being “high risk” of competing risks bias, but very few (6.5%) reported the presence of competing events.

8.1.2 Chapter 3: A review into the presence, reporting, and management of competing events in prediction model development studies.

Given the majority of systematic reviews were considered to be “high risk” in Chapter 2, further research was needed on the management of competing events in prognostic model development studies. Therefore, this chapter aimed to review how prediction model development studies actually handle competing events. The review identified 15 prediction model development studies, producing 25 prediction models in clinical settings where competing events are a likely issue. These prediction models were assessed for the potential for competing risks bias and the management of

competing events. Though the majority (84.0%) of the prediction models were classified as “high risk” of competing risks bias, relatively few (24.0%) utilised statistical methods to appropriately manage the competing events, emphasising that important competing events are often ignored in practice.

8.1.3 Chapter 4: A comparison of time-to-event prognostic models developed using Cox and flexible parametric methods.

To illustrate the application of survival analysis methods without accounting for competing events, this chapter developed two prognostic models to predict the risk of antenatal adverse events in women diagnosed with early-onset pre-eclampsia. The aim was to compare prognostic modelling using Cox proportional hazards regression (Cox, 1972a) and Royston-Parmar flexible parametric regression (Royston and Parmar, 2002). Both models produced almost identical regression coefficient estimates for included predictors. However, the Royston-Parmar model was recommended for use in prognostic model research due to the direct estimation of a flexible and smooth baseline survival function, which allows risk predictions for new individuals, unlike the Cox model which does not estimate the baseline function and so requires non-parametric estimates of the baseline hazard to be incorporated.

8.1.4 Chapter 5: A comparison of prognostic models developed using flexible parametric competing risks methods.

To illustrate the application of survival analysis methods that account for competing risks, this chapter developed two prognostic models which predict the risk of antenatal adverse events in women diagnosed with early-onset pre-eclampsia, while accounting for the competing event of delivery using either cause-specific (Hinchliffe and Lambert, 2013b) or subdistribution (Lambert et al., 2017) approaches within a flexible parametric framework. Both models performed similarly when assessed for predictive performance, however the subdistribution approach resulted in a more

parsimonious model which could be internally validated and adjusted for overfitting, and thus represented important advantages over the cause-specific approach.

8.1.5 Chapter 6: External validation of prognostic models developed using standard and competing risks methods.

In this chapter, the Royston-Parmar (*PREP-RP*) model developed in Chapter 4 and the subdistribution (*PREP-SD*) model developed in Chapter 5 were externally validated using measures adapted to appropriately account for competing events. The *PREP-RP* model, which ignored competing events, performed well at early time points in terms of both calibration and discrimination. However, the calibration of the model was poorer at later time points as the model substantially overestimated the absolute risk of antenatal adverse events. In contrast, the *PREP-SD* model, which accounted for competing events, consistently underestimated the absolute risks of antenatal adverse events. Nevertheless, the model produced risk predictions that better reflect the observed risk of antenatal adverse events at later time-points.

8.1.6 Chapter 7: When are competing risk statistical methods needed in prognostic model research? A simulation study.

Following the findings of the previous chapters, this study used a simulation study to identify how competing risks bias in measures of calibration is affected by the incidence of both the event of interest and competing events. Multiple scenarios were investigated with a range of incidences for the main event of interest and the competing event. Measures of absolute and percentage cumulative risk bias were found to grow as the incidence of the event of interest and the relative proportion of competing events increase. An existing rule-of-thumb, which advises competing events need only be accounted for if the competing event occur at least as often as the event of interest (Berry et al., 2010), was found not to be appropriate for avoiding substantial calibration bias.

8.2 Key findings and recommendations

The findings of the research conducted in this thesis are summarised in the discussion section of each chapter. Four key findings of the thesis are presented in Box 8.1 and are discussed in more detail below:

Box 8.1: Key findings from this thesis

1. Competing events are often present but rarely reported or appropriately managed in prognostic model research.
2. Flexible parametric subdistribution models are recommended to account for competing risks in prognostic model research.
3. Competing risks bias increases with increased incidence of the event of interest and the competing event.
4. Researchers should not use the current rule of thumb for deciding when to use competing risks methods rather than standard survival analysis methods.

8.2.1 Competing events are often present but rarely reported or appropriately managed in prognostic model research.

A consistent finding of the reviews conducted in Chapters 2 & 3 is that in prognostic model research, the presence of competing events is high. In the evaluation of systematic reviews, 90.3% were found to include prediction models with outcomes other than all-cause mortality, and 61.3% were classified as high potential for competing risks bias. In the review of prediction model development studies, mortality was a competing event for 92.0% of the prognostic models, and 73.3% of the prediction model studies were classified as high potential for competing risks bias. These findings are similar to those observed in other systematic reviews; competing risks were found to be present in 78% of reports of randomised controlled trials (Austin and Fine, 2017), 74% of articles published in high impact medical journals (Koller et al., 2012), and 46% of articles which reported Kaplan-Meier curves (Walraven and McAlister, 2016).

Though the presence of competing events was found to be high, the reporting of competing events was low. In the evaluation of systematic reviews, only 6.5% directly

reported the competing events. In the review of prediction model development studies, key terms related to competing events were reported in 20% of the study articles, and the incidence of the competing events was only reported in 26.7% of the study articles. Again these findings appear to be in line with the findings of other systematic reviews; competing risks were discussed as a possible issue in only 18% of articles published in high impact medical journals (Koller et al., 2012), and 34.8% of articles deemed to be susceptible to competing risks bias cited the number of competing events that occurred during the study (Walraven and McAlister, 2016).

When competing events are ignored, this may cause bias in a developed model's absolute risk predictions for the main event of interest. The review of prediction model development studies found that competing events are often not appropriately managed, with only 20% of studies using appropriate methods. Again this reflects the findings of other systematic reviews, with 20% of articles published in high impact journal using appropriate methods (Koller et al., 2012), and 16% of reports of randomised controlled trials (Austin and Fine, 2017).

To the author's knowledge, these are the first reviews that investigate the presence, reporting, and management of competing risks and its associated bias in prediction model research. These reviews aid the understanding of the high presence but rare reporting and appropriate management of competing events; indeed, the findings suggest that a large proportion of prediction model research is susceptible to competing risks bias. However, it has been shown in the later chapters of this thesis (e.g. Chapters 6 and 7) that being susceptible to bias does not always result in biased measures of prognostic performance. The calibration of a prognostic model that predicts an event of interest with low incidence, is developed in the presence of a completing event with a relatively low incidence (compared to the event of interest), or that has a short prediction horizon (resulting in a low occurrence of competing events), is likely to be unaffected by competing risks bias. Additionally, findings from the thesis

suggest the discriminative performance of prognostic models may be hindered by applying competing risks statistical methods (see Table 6.5), though further investigation in a wider range of scenarios is required for a better understanding of this finding. It is argued that the application of competing risks statistical methods to manage competing events whenever they are present would reduce any risk for competing risks bias. However, the requirement for use of competing risks statistical methods in prognostic model research has been shown to depend on the amount of bias expected in measures of prognostic performance, which is influenced by the incidence of both the event of interest and the competing event. The use of these methods may also be influenced by a researcher's decision to use more familiar (standard) time-to-event methods over the more complicated competing risks methods, in order to aid the user's interpretation and understanding of the final prognostic model. Current guidance aimed, at improving the quality of prognostic models (such as the TRIPOD statement), should be updated to specifically mention competing events and encourage researchers to consider the potential for competing risks bias in their prognostic model studies.

The tool developed within these chapters (Figure 2.1) to assess the risk of competing risks bias could be used either when planning the development of, or when assessing, prediction model studies. If a study is considered to be at high risk of competing risks bias, researchers *must* use competing risks statistical methods to appropriately manage the competing events. These competing events should also be better reported; for example, competing risk specific information could be incorporated into updates and extensions of the TRIPOD reporting guidelines to ensure researchers consider and report the presence and proportion of competing events in their studies. Currently the TRIPOD guidance does not mention competing events specifically. It is encouraging that the recent PROBAST (risk of bias assessment tool for prediction model studies) guidance has a specific item on competing risks ("Were complexities in

the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?")

8.2.2 Flexible parametric subdistribution models are recommended to incorporate competing risks in prognostic model research

The benefits of using flexible parametric models, as opposed to parametric or semi-parametric models, are outlined in Chapter 4. The key benefits of using this approach specifically to develop prognostic models are:

1. The direct estimation of a smooth underlying baseline hazard function, and
2. The derivation of absolute risk predictions for individual patients.

The benefits of this approach have been discussed in depth in current literature (Royston and Lambert, 2011), including articles which discuss benefits specifically in terms of prognostic model research (Snell, 2015). Most notably, Professor Sir David Cox, developer of the Cox model, in an interview with Nancy Reid has stated "I would normally want to tackle a problem parametrically… I'm not keen on nonparametric formulations" and "if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically" (Reid, 1994).

Similarly, the benefits of applying the subdistribution, rather than the cause-specific, approach for modelling competing events are discussed in Chapter 5. The key benefits of this approach for developing prognostic models are:

1. Retaining the one-to-one association between the hazard and risk functions, which aids the interpretation of prognostic factors on real world risks,
2. Returning a parsimonious and simpler model, than the cause-specific approach, as only the event of interest needs to be modelled, and
3. The ease of applying standard prognostic model research methods for internal validation, shrinkage, and model updating.

The cause-specific and subdistribution approaches are often compared in the literature (Dignam et al., 2012, Hinchliffe, 2013, Koller et al., 2012, Lau et al., 2009, Putter et al., 2007, Varadhan et al., 2010, Wolkewitz et al., 2014) with some articles again focusing specifically on their use in prognostic model research (Leoce, 2016, Wolbers et al., 2009). It is commonly suggested the two approaches be utilised side-by-side to aid understanding of both the direct and indirect effects of the prognostic factors on the event of interest and competing events (Dignam et al., 2012, Lau et al., 2009, Varadhan et al., 2010, Wolkewitz et al., 2014). However, in prognostic model research these effects are of little importance; instead the key outcome is the accurate estimation of absolute risk predictions. Some authors discuss the convenience of the subdistribution approach, which directly regresses on the cause-specific cumulative incidence function, for answering prognostic modelling research questions (Dignam et al., 2012, Koller et al., 2012).

Given the findings of Chapters 4 & 5, a key recommendation from this thesis is the use of flexible parametric subdistribution models for the development of prognostic models in the presence of competing events. The combination of the two approaches results in the development of parsimonious prognostic models which directly estimate smooth underlying baseline subdistribution hazards that can be modelled without the need to model associations with the competing events.

8.2.3 Competing risks bias increases with increased incidence of the event of interest and the competing event.

The simulation study performed in Chapter 7 demonstrated the strong association between the amount of competing risks bias in measures of calibration and the incidence of both the event of interest and the competing event. Higher levels of competing risks bias were observed in simulation scenarios with lower event of interest incidence and higher competing event incidence (see Figure 7.5). These associations

are similar to those observed by Walraven and team, who found competing risks bias could be determined using the “biased Kaplan-Meier risk estimates and the proportion of all outcomes that were competing events” (van Walraven and Hawken, 2016). A rule of thumb, which advises when the more complex competing risks methods should be applied (Berry et al., 2010), was evaluated in Chapter 7. This rule was found to be insufficient in avoiding substantial bias in cumulative incidence estimates in prediction model studies. An alternative recommendation to evaluate the need for the more complex analysis using a threshold for acceptable percentage bias was proposed. For example, if 10% was considered an acceptable level of percentage bias, then with an event of interest incidence of 5% competing risks should be accounted for if the competing event incidence is at least 17% (Table 7.4). However, to avoid any bias the best approach is to apply competing risks methods whenever competing risks were suspected.

While not appropriately accounting for competing events has been shown to cause bias absolute risk predictions, and thus affects measures of calibration, the inappropriate management of competing events does not appear to have much of an effect of the discriminative performance of the models. In Chapter 6, two models developed to predict the risk of antenatal adverse events in women with early-onset pre-eclampsia were externally validated. Differences between the model’s calibration performance grew as time progressed and more competing events were observed (Figure 6.5), however little difference was observed between the models discriminative performance measures (Table 6.5).

8.3 Discussion points

Many advances in prognostic model and competing risks research have been made throughout the duration of this thesis, while some challenges still remain. These are now summarised.

8.3.1 Current prognostic model research literature

There is a growing interest in prognostic model research, evidenced by the increasing number of scientific articles published on the topic each year (Collins et al., 2015). Many articles investigate the challenges related to the methodology and reporting of prognostic model research (Groenwold et al., 2016, Moons et al., 2009a, Royston and Altman, 2013, Royston et al., 2009, Steyerberg et al., 2013), with some focusing specifically on the presence of competing events (Leoce, 2016, Wolbers et al., 2009). A number of tools have been developed in an attempt to improve the quality of prognostic model research; including the TRIPOD[i] statement, developed to "improve the transparency of the reporting of a prediction models study" (Collins et al., 2015); and the PROBAST[ii] tool, developed to "aid the assessment of the risk of bias and applicability of prediction modelling studies in a systematic review" (Wolff et al., 2019). Additionally, in 2017 a new journal, titled "Diagnostic and Prognostic Research", was created to "ensure that the results of all well-conducted diagnostic and prognostic research are published, regardless of their outcome" (Moons et al., 2017).

A number of systematic reviews highlight the proliferation of prognostic models developed for the same health state to predict similar outcomes, for example 363 prediction models were identified for cardiovascular disease (Damen et al., 2016), 91 for improving the quality of liver resection (Lim et al., 2015), and 63 were identified for functional outcomes post-stroke (Meyer et al., 2015). The large number of available prognostic models could be another barrier which perturbs clinicians from using models to inform clinical decisions (Wessler et al., 2017, Wyatt and Altman, 1995). Further, the development of new prognostic models for the same health states and outcomes as existing models is inefficient, as information captured in previous studies and models is lost (Moons et al., 2012a). Methods for updating and recalibrating

[i] Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.
[ii] Prediction modelling Risk Of Bias ASsessment Tool.

existing models are recommended over the continual development of new prognostic models (Moons et al., 2012a, Royston, 2010). It has been shown that simple recalibration methods are often sufficient to improve the predictive performance of existing prognostic models in new populations or settings (Janssen et al., 2008).

Prognostic models are developed to predict the risk of a future outcome from the intended moment of use (Moons et al., 2015), typically a primary consultation or diagnosis. However, there are many reasons why a patient's risk could change over time; a patient's biomarker measurements may change as the disease progresses, patients may actively alter their behaviour to reduce their risk, and advances in medical treatments could improve risks. Research into the statistical methods to incorporate the change of individual risks over time into prognostic models is advancing. Dynamic prediction models retain prognostic accuracy by "evolving over time in response to observed changes" (Jenkins et al., 2018). Models are updated, either at discrete time-points as populations, treatments, and clinical practices change, or continually over time. Landmark prediction models allow predictions to be made from different time points, for example at diagnosis, then again three years after diagnosis (Lambie et al., 2019). Landmark models update expected risk predictions over time, allowing clinicians and patients to reassess risks at different time points. Finally, joint modelling methods can be applied to prognostic models to allow repeated prognostic factor measurements to be incorporated into the risk predictions (Proust-Lima and Taylor, 2009). This can aid prediction when a prognostic factor measurement is highly variable within individuals (such as blood pressure), or when risk is associated with the trajectory (pattern over time) of the prognostic factor.

The prognostic model literature supports the investigation of a single specific outcome (Steyerberg et al., 2013). However, in clinical practice multiple outcomes may be of interest. The multistate framework was introduced for standard time-to-event and competing risks analysis in Chapter 1. This framework can be extended to incorporate

numerous events of interest as well as intermediate events, such as recovery or adverse events, as depicted in Figure 8.1. Multi-state prognostic models allow the prediction of multiple outcomes of interest and help describe the patients transitions between different health states (Javanmardi et al., 2018), providing a broader view of future disease progression.

Figure 8.1: Complex multi-state model for prognosis following bone marrow transplantation in leukaemia patients

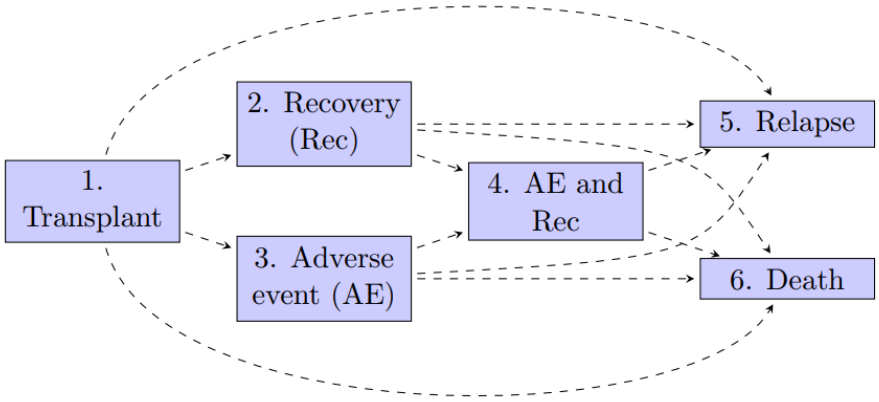


Figure 1: A six-states model for leukemia patients after bone marrow transplantation

8.3.2 Current competing risks literature

It has been shown that competing events are rarely utilised in prognostic model research. There are increasing efforts to make competing risks methods more accessible to researchers. A number of user friendly statistical software packages have been developed in Stata to enable competing risks analysis, including the *stpm2* function (Lambert and Royston, 2009), an extension *stpm2cif* (Hinchliffe and Lambert, 2013a), and the *multistate* function (Crowther and Lambert, 2019). Multiple online tools have been developed for use as teaching tools and to aid understanding of key methodological concepts, including an interactive graph which demonstrates the link between cause-specific hazards and the cumulative incidence function (Lambert) and an interactive cancer survival prediction tool (Mozumder). Earlier tutorials on

competing risks analysis focused on applications in cancer progression (Fine and Gray, 1999, Hinchliffe, 2013, Pintilie, 2007). However, recently there has been an increase in discussions of the use of these methods for other health states, including coronary heart disease (Wolbers et al., 2009), nephrology (Noordzij et al., 2013), and pneumonia (Wolkewitz et al., 2014). All of the aforementioned resources aim to increase the accessibility of competing risks methods to researchers, in an attempt to increase the use of the methods in practice.

The more popular Fine and Gray model (Fine and Gray, 1999) is not discussed in detail during this thesis, as the focus is on the use of flexible parametric competing risks approaches. In brief, the Fine and Gray model is a semi-parametric model on the subdistribution hazards scale. As such, the model does not directly estimate the underlying cumulative baseline subdistribution hazard function and does not provide absolute risk predictions. However, the model may be useful in prognostic model research for identifying associations between prognostic factors and the real-world risk of the event of interest, accounting for the competing event.

The statistical methods which allow for adaptive prognostic model predictions, discussed previously, are also being extended to incorporate competing events. The landmark model approach has been extended to allow dynamic prediction of competing risks while incorporating time-dependant covariates (Nicolaie et al., 2013), allowing prediction to be made at specified timepoints after initial diagnosis. Advances to joint modelling methods have been made to model multiple longitudinal measurements with competing risks time-to-event outcomes (Andrinopoulou et al., 2017), producing dynamically updated absolute risk predictions.

8.4 Limitations and further research

Some limitations of the work contained within this thesis are outlined below, alongside suggestions of further research that will address these limitations.

Neither of the reviews conducted in Chapters 2 and 3 encompass the entirety of the prediction model research literature. These reviews were performed to provide a snapshot of the literature in order to assess the presence, management, and reporting of competing events in prediction model research. As such, many systematic reviews of prediction model development studies, and prediction model development study articles are not considered. To the authors knowledge, these are the largest reviews to assess competing risks bias in prediction model literature that have been conducted. The finding of high competing risk incidence from the evaluation of systematic reviews of prediction model development studies (Chapter 2) is likely to be generalisable, as most of the included systematic reviews searched for articles from database inception and prediction model outcomes are unlikely to have changed over time. The prediction model development studies identified for inclusion in the review conducted in Chapter 3 were purposefully selected from clinical settings where competing events are likely to be an issue. The finding of high incidence of competing events in these articles is thus likely to be inflated compared to the incidence in all prediction model research. Further research could extend both of the reviews conducted in this thesis to incorporate all prediction model literature, including articles published since the review was conducted (April 2017). This would allow for a better understanding of the presence of competing risks in all prognostic model research, and could perhaps identify whether there were trends in the presence, reporting, and management of competing events over different health states, or indeed over time.

The presence of competing risks bias in the studies included in Chapters 2 & 3 was not known, and thus was evaluated using a risk tool which assessed three criteria (Figure 2.1). It was considered out of the scope of the thesis to obtain individual participant data from each study to thoroughly assess the presence of competing events. The risk tool simply combines factors which were thought to be applicable to competing risks bias in prediction model studies, and which were likely to be reported.

However, this tool has not been evaluated. Further research could go on to evaluate the risk tool in prediction model studies to identify how accurate it is at identifying the presence of competing events. If found to be accurate, the tool could be utilised to inform the study design and statistical methods used in future prediction model development studies.

The prognostic models developed for predicting outcomes in early-onset pre-eclampsia patients (Chapters 4 & 5) were potentially underpowered. The PREP study was designed and powered to predict the risk of maternal complication, including delivery before 34 weeks, of which there were 633 events (Thangaratinam et al., 2017). In this thesis, the alternative outcome of antenatal adverse events was investigated, either ignoring or incorporating delivery as a competing event, of which there were only 75 events. This was a pragmatic use of available data, and was considered a sufficient sample size to indicate differences in semi-parametric and flexible parametric, and standard time-to-event and competing risks, approaches. Furthermore, the small number of antenatal adverse events restricted the analysis to the composite outcome; there were too few events to examine individual components separately. The adverse events differ in severity and would be managed differently, thus the evaluation of each type of adverse event would be more clinically meaningful but would require a much larger study. External validation in other settings could be explored to determine the generalisability of the competing risks model. Finally, the adverse events which occurred following delivery of the baby were not considered during this thesis, as the focus was on competing events. Further research into early-onset pre-eclampsia prognosis would include conducting a larger study, powered to reduce the potential for overfitting, and thus to develop a more robust prognostic model for adverse events. If the sample size were sufficiently large, this study could investigate the risks for each different type of adverse event separately, or indeed could use multi-state modelling methods to develop a prognostic model for transitions between antenatal adverse

events, delivery, and postnatal adverse events. Such a model would be beneficial to clinicians and patients and would facilitate informed treatment decision making for women diagnosed with early-onset pre-eclampsia.

There is currently little research into methods for validation or recalibration of prognostic models developed using competing risks methods. Recent research has focused on developing calibration curves for competing risks models (Gerds et al., 2014) and adapting concordance for competing events (Wolbers et al., 2014). Further methodological research into measures of calibration and discrimination, adapted for the presence of competing events is required. Additionally, standard prognostic model updating and recalibration methods are not easily transferable to cause-specific competing risks models, and further methodological research is required.

The simulation studies conducted in Chapter 7 utilise simple constant hazard functions (exponential models), which are usually not considered to be biologically plausible in medical research (Crowther and Lambert, 2013). The studies also focused on parametric absolute risk estimates, and failed to incorporate prognostic factor associations. To reflect prognostic model research the simulations should have incorporated the effects of prognostic factor associations with both the event of interest and the competing event, and assessed absolute risk predictions from flexible parametric regression equations against expected risks. Further simulation studies which incorporate more complex and realistic hazard functions, modelled using flexible parametric modelling of the underlying hazard functions, are planned for future research. This future research will also investigate the impact of incorporating prognostic factor associations with the event of interest and the competing event, and evaluate bias in both measures of calibration and discrimination. This research will provide a greater understanding of the impact of event of interest and competing event incidences, and associations between prognostic factors and the event of interest or competing event, on the biases caused by mismanaging competing events.

8.5 Final conclusions

This thesis has evaluated the use of statistical methods for developing and validating competing risks prognostic models for use in clinical practice, including an applied example in early-onset pre-eclampsia prognosis. It provides new empirical evidence that competing events are often ignored in prognostic model research, including development studies and systematic reviews. It also highlights the potential for large bias in risk predictions when ignoring competing risks methods. While methods for developing competing risks prognostic models are advanced and widely accessible, further methodological research is required to ensure models can be properly validated and updated.

9 APPENDIX I

Data Extraction Sheet:

Evaluation of Systematic Reviews of Prediction Model Studies

Version 3.0

Item 1: What were the characteristics of each systematic review?	
Title of systematic review article:	
Endnote reference:	
Date restrictions reported in search strategy:	**Start date:** **End date:**
Primary aim of systematic review*:	
Aims to identify prediction models for a specific outcome? ☐ YES ☐ NO ☐ UNCLEAR	
Aims to identify prediction models in a specific population? ☐ YES ☐ NO ☐ UNCLEAR	
Number of prediction models (& studies) identified by systematic review¥	

*The primary aim is taken from systematic review abstract.

¥Some systematic reviews may identify a study article which describes the development of multiple prediction models, additionally some articles may identify both development and validation articles, thus the number of models and studies may not be equal.

Item 2: What is the potential for competing risks bias affecting each systematic review?		
Criterion 1: The prediction model investigates outcomes other than all-cause mortality		
Outcomes of the prediction models identified by the review	All-cause mortality or composite ☐	
	Single (not ACM / PFS) ☐	
Criterion 2: The baseline population contains frail and/or elderly populations		
Disease and health states of prediction model populations		
Population age at baseline	Measures:	Results:
Criterion 3: The prediction horizon is sufficiently long to enable competing events to occur		
Prediction horizon of prediction models identified by the review*	Measures:	Results:
Assessing the potential for competing risks bias within each systematic review		
Criterion 1: The review contains prediction models with outcomes other than all-cause mortality. *If no select None, if yes proceed to Criterion 2:*	None ☐	
Criterion 2: The review contains prediction models developed in elderly or frail populations. *If no select Low, if yes proceed to Criterion 3:*	Low ☐	
Criterion 3: The review contains prediction models which predict outcomes after 1 year or more? *If no select Moderate, if yes select High.*	Moderate ☐	
	High ☐	

*If specific prediction time horizon not reported, report measure of follow-up times

Item 3: Were competing events reported in the systematic review article?	
Term “competing risk(s)” or “competing event(s)” used?	☐ YES ☐ NO
If yes copy text:	
Other competing risks terms used in the article? Such as: *competing cause(s); competing bias; cause-specific; subdistribution; cumulative incidence; Fine and Gray;*	☐ YES ☐ NO
If yes copy text:	
Kaplan-Meier estimates or curves presented or discussed?	☐ YES ☐ NO
If yes copy text:	

Item 4: Was competing risks part of the quality assessment performed within the systematic review?	
Quality assessment performed?	☐ YES ☐ NO
Tool used:	

10 APPENDIX II

Table 10.1: Characteristics of systematic reviews included in evaluation of systematic reviews of prediction model studies (Chapter 2)

Systematic review reference	Publication title	Search dates of the review	Primary outcome(s) specified in the review?	Primary population specified in the review?	No. of models (primary study articles) in each review
(Ayerbe et al., 2016)	Clinical assessment of patients with chest pain; a systematic review of predictive tools	Inception to July 2015	Yes	Yes	13 (12)
(Brunelli and Prefumo, 2015)	Quality of first trimester risk prediction models for pre-eclampsia: a systematic review	Inception to July 2013	Yes	Yes	38 (24)
(Caragata et al., 2016)	Acute kidney injury following liver transplantation: a systematic review of published predictive models	Inception to May 2015	Yes	Yes	7 (7)
(Damen et al., 2016)	Prediction models for cardiovascular disease risk in the general population: systematic review	Inception to June 2013	Yes	Yes	363 (125)
(Echouffo-Tcheugui et al., 2015)	Population risk prediction models for incident heart failure: a systematic review	January 1990 to August 2014	Yes	Yes	28 (13)
(Ensor et al., 2016)	Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation	Inception to June 2014	Yes	Yes	3(3)
(Gray et al., 2016)	Risk Prediction Models for Lung Cancer: A Systematic Review	1985 to July 2014	Yes	No	16 (25)
(Haskins et al., 2015)	Validation and impact analysis of prognostic clinical prediction rules for low back pain is needed: a systematic review	Inception to July 2013	No	Yes	30 (35)

Systematic review reference	Publication title	Search dates of the review	Primary outcome(s) specified in the review?	Primary population specified in the review?	No. of models (primary study articles) in each review
(Hilkens et al., 2016)	Prediction models for intracranial hemorrhage or major bleeding in patients on antiplatelet therapy: a systematic review and external validation study	Inception to December 2014	Yes	Yes	5 (5)
(Kohn et al., 2015)	Prognostic Accuracy of Clinical Prediction Rules for Early Post-Pulmonary Embolism All-Cause Mortality: A Bivariate Meta-analysis	January 2000 to March 2014	Yes	Yes	11 (40)
(Lim et al., 2015)	Improving the quality of liver resection: a systematic review and critical analysis of the available prognostic models	May 1999 to March 2012	No	Yes	91 (91)
(Linsell et al., 2016)	Prognostic Factors for Behavioral Problems and Psychiatric Disorders in Children Born Very Preterm or Very Low Birth Weight: A Systematic Review	January 1990 to June 2014	No	Yes	30 (15)
(Luo et al., 2015)	A systematic review of predictive models for asthma development in children	Inception to June 2015	Yes	Yes	30 (32)
(Mahajerin et al., 2015)	Hospital-associated venous thromboembolism in pediatrics: a systematic review and meta-analysis of risk factors and risk-assessment models	Inception to May 2014	Yes	Yes	3 (60)
(Mao et al., 2015)	Predictors associated with stroke after coronary artery bypass grafting: A systematic review	January 1990 to September 2014	Yes	Yes	14 (14)
(Marques et al., 2015)	The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis	2003 to September 2014	Yes	Yes	13 (45)
(Marufu et al., 2015)	Risk scoring models for predicting perioperative morbidity and mortality in people with fragility hip fractures: Qualitative systematic review	1966 to April 2015	Yes	Yes	25 (29)
(Meyer et al., 2015)	A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation	Inception to January 2013	Yes	Yes	63 (27)

Systematic review reference	Publication title	Search dates of the review	Primary outcome(s) specified in the review?	Primary population specified in the review?	No. of models (primary study articles) in each review
(O'Caoimh et al., 2015)	Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults	1965 to November 2014	No	Yes	23 (46)
(Oliver et al., 2015)	Risk assessment tools validated for patients undergoing emergency laparotomy: a systematic review	1980 to November 2014	No	Yes	25 (20)
(Quinlivan et al., 2016)	Which are the most useful scales for predicting repeat self-harm? A systematic review evaluating risk scales using measures of diagnostic accuracy	Inception to February 2015	Yes	Yes	11 (8)
(Salz et al., 2015)	Are we ready to predict late effects? A systematic review of clinically useful prediction models	Inception to April 2014	No	Yes	16 (14)
(Silver et al., 2015)	Risk prediction models for contrast induced nephropathy: systematic review	Inception to March 2015	Yes	Yes	12 (16)
(Silverberg et al., 2015)	Systematic review of multivariable prognostic models for mild traumatic brain injury	1970 to June 2013	No	Yes	49 (26)
(Smit et al., 2015)	Childhood asthma prediction models: a systematic review	Inception to June 2014	Yes	Yes	12 (12)
(Tang et al., 2015)	Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review	January 2009 to March 2014	Yes	No	Unclear* (21)
(Usher-Smith et al., 2016)	Risk Prediction Models for Colorectal Cancer: A Systematic Review	January 2000 to March 2014	Yes	Yes	52 (40)
(Walsh et al., 2016)	Systematic review of risk prediction models for falls after stroke.	Inception to June 2015	Yes	Yes	18 (12)
(Warnell et al., 2015)	Predicting perioperative mortality after oesophagectomy: a systematic review of performance and methods of multivariate models	1990 to 2012	Yes	Yes	11 (20)

Systematic review reference	Publication title	Search dates of the review	Primary outcome(s) specified in the review?	Primary population specified in the review?	No. of models (primary study articles) in each review
(Williams et al., 2016)	Risk prediction models for colorectal cancer in people with symptoms: a systematic review	January 2000 to March 2014	Yes	Yes	15 (18)
(Wilson et al., 2016)	Risk prediction models for acute kidney injury following major noncardiac surgery: systematic review	Inception to June 2014	Yes	Yes	7 (6)

***Multiple models listed multiple times with different cut points for assessment**

11 APPENDIX III

Data Extraction Sheet:

Review of Prediction Model Development Studies

Version 3.0

Item 1: What were the characteristics of each prediction model development study?		
Title of prediction model study:		
Endnote reference:		
Systematic review reference:		
Number of individual prediction models developed in prediction model study:		
Source of study data:	Cohort ☐	RCT ☐
	Case-control ☐	Other (list below) ☐
Number of participants*:	Development	
	Validation	
	Total	

*if multiple models are reported in the article then report the greatest number of eligible participants

Item 2: What is the potential for competing risks bias affecting each prediction model?			
Criterion 1: The prediction model investigates outcomes other than all-cause mortality			
Outcomes of the prediction model	All-cause mortality or composite ☐		
	Single (not ACM / PFS) ☐		
Criterion 2: The baseline population contains frail and/or elderly populations			
Disease and health states of prediction model population			
Population age at baseline	Measures:	Results:	
Criterion 3: The prediction horizon is sufficiently long to enable competing events to occur			
Prediction horizon of prediction model *	Measures:	Results:	
Assessing the potential for competing risks bias within each prediction model			
Criterion 1: The prediction model investigates outcomes other than all-cause mortality. *If no select None, if yes proceed to Criterion 2:*	None ☐		
Criterion 2: The prediction model was developed in elderly or frail populations. *If no select Low, if yes proceed to Criterion 3:*	Low ☐		
Criterion 3: The prediction models predict outcomes after 1 year or more? *If no select Moderate, if yes select High.*	Moderate ☐		
	High ☐		
Potential competing events			
Potential competing events likely to prevent the prediction model outcome from occurring:			

*If specific prediction horizon not reported, report measure of follow-up time

Item 3: Were competing events reported in the published prediction model study articles?		
Reported number (%) of events:	Prediction model events:	Competing events:
	Total observed events:	Proportion competing events:
Term “competing risk(s)” or “competing event(s)” used? ☐ YES ☐ NO		
If yes copy text:		
Other competing risks terms used in the article? ☐ YES ☐ NO Such as: *competing cause(s); competing bias; cause-specific; subdistribution; cumulative incidence; Fine and Gray;*		
If yes copy text:		
Kaplan-Meier estimates or curves presented or discussed? ☐ YES ☐ NO		
Number of prognostic factors in final prediction models:		
Prognostic factors associated with competing events:	Age ☐	Comorbidities ☐

Item 4: How were competing events managed in each prediction model study?		
Competing risks statistical methodology used?	☐ YES ☐ NO	
Competing events censored?	☐ YES ☐ NO	
Competing events excluded?	☐ YES ☐ NO	
How was the model validated?	None ☐	Apparent validation ☐
	Internal validation ☐	External validation ☐
How were competing events managed?		

12 APPENDIX IV

Table 12.1: Reasons for inclusion/exclusion of systematic reviews (from chapter 2) likely to contain models affected by competing events

Systematic review reference	Potential for competing risks bias	Baseline populations	Corresponding characteristic (Koller et al., 2012)	Include/exclude and reason
(Kohn et al., 2015)	None	Acute pulmonary embolism		Exclude No potential
(Oliver et al., 2015)	None	Emergency laparotomy		Exclude No potential
(Warnell et al., 2015)	None	Oesophagectomy for cancer in adults		Exclude No potential
(Luo et al., 2015)	Low	Children		Exclude Low potential
(Smit et al., 2015)	Low	Children with asthma-like symptoms		Exclude Low potential
(Caragata et al., 2016)	Moderate	Liver transplantation	Transplant	Exclude Moderate potential
(Lim et al., 2015)	Moderate	Patients undergoing liver resection		Exclude Moderate potential
(Mao et al., 2015)	Moderate	Coronary artery bypass grafting	Coronary heart disease	Exclude Moderate potential
(Meyer et al., 2015)	Moderate	Patients receiving post-stroke inpatient rehabillitation	Stroke	Exclude Moderate potential
(Silver et al., 2015)	Moderate	Undergoing procedure using iodinated radiocontrast		Exclude Moderate potential
(Brunelli and Prefumo, 2015)	Moderate	First trimester pregnancy		Exclude Moderate potential
(Mahajerin et al., 2015)	Moderate	Pediactric hospital patients		Exclude Moderate potential
(Ayerbe et al., 2016)	High	Recent onset chest pain	Cardiac failure	Include
(Damen et al., 2016)	High	General population		Exclude Not susceptible

Systematic review reference	Potential for competing risks bias	Baseline populations	Corresponding characteristic (Koller et al., 2012)	Include/exclude and reason
(Echouffo-Tcheugui et al., 2015)	High	General population		Exclude Not susceptible
(Ensor et al., 2016)	High	Cessation of treatment venous thrombembolism		Exclude Not susceptible
(Gray et al., 2016)	High	General population		Exclude Not susceptible
(Haskins et al., 2015)	High	Low back pain		Exclude Not susceptible
(Hilkens et al., 2016)	High	Patients on antiplatelet therapy	Cardiac failure	Include
(Linsell et al., 2016)	High	Very preterm or very low birth weight infants		Exclude Not susceptible
(Marques et al., 2015)	High	General population		Exclude Not susceptible
(Marufu et al., 2015)	High	Hip fracture operation		Exclude Not susceptible
(O'Caoimh et al., 2015)	High	Community-dwelling older adults	Elderly patients	Include
(Quinlivan et al., 2016)	High	Presenting with self-harm or attempted suicide		Exclude Not susceptible
(Salz et al., 2015)	High	Completing treatment for cancer	Prostate cancer, colorectal cancer, breast cancer	Include
(Silverberg et al., 2015)	High	Mild traumatic brain injury		Exclude Not susceptible
(Tang et al., 2015)	High	Varied, including general population, elderly, and with diabetes		Exclude Not susceptible
(Usher-Smith et al., 2016)	High	General population		Exclude Not susceptible
(Walsh et al., 2016)	High	Stroke	Stroke	Include
(Williams et al., 2016)	High	Symptomatic of colorectal cancer	Colorectal cancer	Include
(Wilson et al., 2016)	High	Major non-cardiac surgery		Exclude Not susceptible

13 APPENDIX V

Full list of published prediction model study articles identified from six systematic review articles, including reasons for inclusion/exclusion.

Systematic review	Articles	Include/Exclude	Reason
(Ayerbe et al., 2016)	Bassan 2004	Exclude	Neural tree – machine learning
	Bjork 2006	Exclude	Logistic regression
	Bosner 2010	Exclude	Logistic regression
	Gencer 2010	Exclude	Logistic regression
	Genders 2012	Exclude	Logistic regression
	Goodacre 2002	Exclude	Logistic regression
	Gruseels 1995	Exclude	Logistic regression
	Pryor 1993	Exclude	Model validation study
	Sanchez 2007	Exclude	Logistic regression
	Sekhri 2008	**Include**	**Cox regression**
	Sox 1990	Exclude	Logistic regression
	Tierney 1985	Exclude	Logistic regression
(Hilkens et al., 2016)	**Ariesen 2006**	**Include**	**Cox regression**
	Barra 2013	Exclude	Risk score: derrivation of weights not through regression methods

Systematic review	Articles	Include/Exclude	Reason
	Cuschieri 2014	**Include**	**Cox regression**
	Ducrocq 2010	Exclude	Logistic regression
	Geraghty 2012	Exclude	Conference abstract – insufficient information
(O'Caoimh et al., 2015)*	Alessi 2003	Exclude	Risk score: derrivation of weights not through regression methods
	Boult 1993	Exclude	Logistic regression
	Carey 2004	Exclude	Logistic regression
	Carey 2008	**Include**	**Cox regression**
	Covinsky 2006	Exclude	Logistic regression
	Crane 2010	Exclude	Logistic regression
	Dalby 1990	Exclude	Risk score: derrivation of weights not through regression methods
	Damush 2004	Exclude	Logistic regression
	Fitzgerald 2015	Exclude	Conference abstract – insufficient information
	Freedman 1996	Exclude	Logistic regression
	Han 2012	Exclude	Logistic regression
	Herbert 1996	Exclude	Logistic regression
	Kerse 2008	Exclude	Risk score: derrivation of weights not through regression methods
	Lee 2006	Exclude	Logistic regression
	Lyon 2007	Excluded	Logistic regression
	Mazzaglia 2007	Exclude	Logistic regression
	Reuben 2002	Exclude	Logistic regression

Systematic review	Articles	Include/Exclude	Reason
	Roos 1988	Exclude	Logistic regression
	Saliba 2001	Exclude	Logistic regression
	Schoenberg 2009	**Include**	**Cox regression**
	Shelton 2000	Exclude	Logistic regression
	St John 2014	Exclude	Risk score: derrivation of weights not through regression methods
	Suijker 2014	Exclude	Logistic regression
(Salz et al., 2015)	Alemozaffar 2011	Exclude	Logistic regression
	Bevilacqua 2012	**Include**	**Cox regression**
	Briganti 2010	**Include**	**Cox regression**
	Chipman 2014	Exclude	Generalized estimating equations
	Eastham 2008	Exclude	Generalized estimating equations
	Ezaz 2014	**Include**	**Cox regression**
	Gallina 2008	Exclude	Conference abstract – insufficient information
	Ganz 1993	Exclude	Logistic regression
	Kilminster 2011	Exclude	Logistic regression
	Kovalchik 2013	**Include**	**Cox regression**
	Langendijk 2009	Exclude	Logistic regression
	Mathieu 2013	**Include**	**Cox regression**
	Romond 2012	**Include**	**Parametric and Cox regression**

Systematic review	Articles	Include/Exclude	Reason
	Travis 2005	**Include**	**Time-to-event methods**
(Walsh et al., 2016)	Ashburn 2008	Exclude	Logistic regression
	Baetens 2011	Exclude	Logistic regression
	Chen 2015	Exclude	Poisson regression
	Kerse 2008	Exclude	Logistic regression
	Mackintosh 2006	Exclude	Logistic regression
	Nakagawa 2008	**Include**	**Cox regression**
	Nyberg 1997	**Include**	**Cox regression**
	Olsson 2005	Exclude	Conference abstract – insufficient information
	Rabadi 2008	Exclude	Logistic regression
	Rapport 1993	Exclude	Risk score: derrivation of weights not through regression methods
	Sze 2001	Exclude	Logistic regression
	Tilson 2012	Exclude	Classification and regression tree – Machine learning
(Williams et al., 2016)	Adelstein 2010	Exclude	Logistic regression
	Adelstein 2011	Exclude	Logistic regression
	Ballal 2010	Exclude	Model validation study
	Collins 2012	Exclude	Model validation study
	Fijten 1995	Exclude	Logistic regression
	Hamilton 2005	Exclude	Logistic regression
	Hamilton 2009	Exclude	Logistic regression

Systematic review	Articles	Include/Exclude	Reason
	Hamilton 2013	Exclude	Model validation study
	Hippisley-Cox 2012	Include	Cox regression
	Hippisley-Cox 2013 (1)	Exclude	Logistic regression
	Hippisley-Cox 2013 (2)	Exclude	Logistic regression
	Hodder 2004	Exclude	Model validation study
	Hurst 2007	Exclude	Logistic regression
	Lam 2002	Exclude	Risk score: derrivation of weights not through regression methods
	Mahadavan 2012	Exclude	Logistic regression
	Marshall 2011	Exclude	Model validation study
	Rai 2008	Exclude	Model validation study
	Selvachandran 2002	Exclude	Model validation study

***O'Caoimh found a further 23 papers eligible for the systematic review; However as the published article focused on the unique instuments, and these articles described the external validation of the instruments, it is hard to identify these from the published article and thus they are not reported in this table.**

14 APPENDIX VI

Table 14.1: Characteristics of prediction model studies included in review of prediction model development studies (Chapter 3)

Systematic review reference	Prediction model study reference	Prediction model study article title	Number of individual models developed[1]	Source of study data[2]	Number of participants[3]: Development Validation Total
(Ayerbe et al., 2016)	**(Sekhri et al., 2008)**	Incremental prognostic value of the exercise electrocardiogram in the initial assessment of patients with suspected angina: cohort study	3	Cohort	8,176 N/A 8,176
(Hilkens et al., 2016)	**(Ariesen et al., 2006)**	Predictors of risk of intracerebral haemorrhage in patients with a history of TIA or minor ischaemic stroke	1	Cohort	12,648 12,648 12,648
	(Cuschieri et al., 2014)	Risk factors for acute gastrointestinal bleeding following myocardial infarction in veteran patients who are prescribed clopidogrel	1	Cohort	3,218 3,218 3,218
(O'Caoimh et al., 2015)	**(Carey et al., 2008)**	Prediction of Mortality in Community-Living Frail Elderly People with Long-Term Care Needs	1	Cohort	2,232 1,667 3,899
	(Schonberg et al., 2009)	Index to Predict 5-Year Mortality of Community-Dwelling Adults Aged 65 and Older Using Data from the National Health Interview Survey	1	Cohort	16,077 8,038 24,115
(Salz et al., 2015)	**(Bevilacqua et al., 2012)**	Nomograms for Predicting the Risk of Arm Lymphedema after Axillary Dissection in Breast Cancer	3	Cohort	1,054 1,054 1,054

Systematic review reference	Prediction model study reference	Prediction model study article title	Number of individual models developed[1]	Source of study data[2]	Number of participants[3]: Development Validation Total
	(Briganti et al., 2010)	Predicting Erectile Function Recovery after Bilateral Nerve Sparing Radical Prostatectomy: A Proposal of a Novel Preoperative Risk Stratification	1	Cohort	435 435 435
	(Ezaz et al., 2014)	Risk Prediction Model for Heart Failure and Cardiomyopathy After Adjuvant Trastuzumab Therapy for Breast Cancer	1	Cohort	832 832 1,664
	(Kovalchik et al., 2012)	Absolute Risk Prediction of Second Primary Thyroid Cancer Among 5-Year Survivors of Childhood Cancer	3	Nested case-control	12,150 2,966 15,116
	(Mathieu et al., 2014)	Nomograms to predict late urinary toxicity after prostate cancer radiotherapy	3	Cohort & RCT	965 N/A 965
	(Romond et al., 2012)	Seven-Year Follow-Up Assessment of Cardiac Function in NSABP B-31, a Randomized Trial Comparing Doxorubicin and Cyclophosphamide Followed by Paclitaxel (ACP) With ACP Plus Trastuzumab As Adjuvant Therapy for Patients With Node-Positive, Human Epidermal Growth Factor Receptor 2–Positive Breast Cancer	1	RCT	1,690 1,690 1,690
	(Travis et al., 2005)	Cumulative Absolute Breast Cancer Risk for Young Women Treated for Hodgkin Lymphoma	2	Nested case-control	370 92 462
(Walsh et al., 2016)	**(Nakagawa et al., 2008)**	Development of an assessment sheet for fall prediction in stroke inpatients in convalescent rehabilitation wards in Japan	1	Cohort	704 N/A 704

Systematic review reference	Prediction model study reference	Prediction model study article title	Number of individual models developed[1]	Source of study data[2]	Number of participants[3]: Development Validation Total
	(Nyberg and Gustafson, 1997)	Fall Prediction Index for Patients in Stroke Rehabilitation	1	Cohort	135 N/A 135
(Williams et al., 2016)	**(Hippisley-Cox and Coupland, 2012)**	Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm	2	Cohort	2,351,052 1,236,601 3,587,653

[1]Number of individual prediction models developed within prediction model study
[2]Source of model development data
[3]The number of participants included in prediction model study, including those for model development (top), for validation (middle), and total number of participants (bottom). Some participants were used for both model development and validation (internal validation with bootstrapping). Where multiple individual prediction models were developed in the same prediction model study, the maximum number of participants used to develop or validate a model is reported.
TIA=Transient ischaemic attack, NSABP=National Surgical Adjuvant Breast and Bowel Project, RCT= Randomised controlled trial

15 APPENDIX VII

Table 15.1: Potential competing events for each prediction model study included in the review of prediction model development studies (Chapter 3)

Prediction model study reference	Prediction model outcomes	Disease and health state of prediction model population	Potential competing events
(Sekhri et al., 2008)	Death due to coronary heart disease or non-fatal acute coronary syndrome	Suspected angina	Other mortality, preventative treatments.
(Ariesen et al., 2006)	Intracerebral haemorrhage	Patients with ischaemic stroke or transient ischemic attack	Mortality
(Cuschieri et al., 2014)	Acute gastrointestinal (GI) bleeding	Patients with myocardial infarction & prescribed clopidogrel	Mortality
(Carey et al., 2008)	All-cause mortality	Community living frail elderly people with long-term care needs	
(Schonberg et al., 2009)	5-year mortality	Community dwelling adults aged 65 and older	
(Bevilacqua et al., 2012)	Lymphedema	Axillary lymph node dissection in breast cancer	Mortality, breast cancer recurrence, other cancer.
(Briganti et al., 2010)	Erectile function recovery	Prostate cancer treated with bilateral nerve sparing prostsectomy	Mortality, prostate cancer recurrence, other cancer.
(Ezaz et al., 2014)	Heart failure and cardiomyopathy	Patients receiving adjuvant trastuzumab therapy for early-stage breast cancer	Mortality, breast cancer recurrence, other cancer.
(Kovalchik et al., 2012)	Second primary thyroid cancer	Survivors of childhood cancers	Mortality, complete removal of thyroid gland, other cancer.

Prediction model study reference	Prediction model outcomes	Disease and health state of prediction model population	Potential competing events
(Mathieu et al., 2014)	Urinary toxicity, urinary frequency, dysuria	Prostate cancer radiotherapy	Mortality, prostate cancer recurrence, other cancer.
(Romond et al., 2012)	Severe congestive heart failure or cardiac death	Patients with node-positive breast cancer	Mortality, breast cancer recurrence, other cancer.
(Travis et al., 2005)	Breast cancer	Young women treated for Hodgkin's lymphoma	Mortality, Hodgkin's lymphoma recurrence, other cancer.
(Nakagawa et al., 2008)	Falls	Stroke inpatients in convalescent rehabilitation wards	Mortality, immobility, hospital discharge
(Nyberg and Gustafson, 1997)	Falls	Patients in stroke rehabilitation	Mortality, immobility, hospital discharge
(Hippisley-Cox and Coupland, 2012)	Colorectal cancer	Patients with suspected colorectal cancer	Mortality, other cancer.

16 APPENDIX VIII

Instances in which competing risks terms were used within published prediction model study articles.

(Kovalchik et al., 2012)
Absolute risk is the probability that an individual with a specific risk profile will develop disease by a given age in the presence of ***competing events***. ***Competing risks. Competing events*** for [second primary thyroid cancer] were death, self-reported complete removal of the thyroid gland, and other second primary cancers, which were determined from pathology reports with follow-up to January 1, 2010. Estimates of absolute risk combined semiparametric estimates of baseline incidences and [relative risks] for [second primary thyroid cancer] and ***competing risk*** comprising death, thyroid removal, or other second primary cancers. [Relative risks] for [second primary thyroid cancer] were estimated from the pooled cohort and case-control studies, but ***competing event*** [relative risks] were estimated from the [Childhood Cancer Survivor Study] cohort only. Hazard models for [second primary thyroid cancer] under [model 1] and [model 2] and ***competing event*** models for [model 1], [model 2], and [model 3] followed a Cox proportional hazards model. An [first primary cancer] diagnosis of Hodgkin lymphoma ([relative risk], 1.7; 95% CI, 1.6 to 1.9) and any prior diagnosis of a thyroid nodule ([relative risk], 1.7; 95% CI, 1.3 to 2.2) were also found to be significant prognostic variables for ***competing events*** in the self-report-only model. In the ***competing events*** model, the presence of treatment variables with large [relative risk] —1.6 (95% CI, 1.5 to 1.7) for alkylating agents (yes/no), 2.1 (95% CI, 1.8 to 2.3) for radiation (yes/no), and 1.8 (95% CI, 1.6 to 1.9) for radiation to the neck (yes/no)—resulted in an attenuation of the [relative risk] associated with thyroid nodules ([relative risk], 1.4; 95% CI, 1.1 to 1.8). All treatment-related factors were strongly associated with the ***competing risk*** model for [model 3]; the risk association for thyroid nodules ([relative risk], 1.5; 95% CI, 1.1 to 1.9) was unchanged from [model 2].

(Romond et al., 2012)

For this purpose, we have developed a prediction model for ***cumulative incidence*** of cardiac events based on identified pretreatment risk factors.

Patients who crossed over from [Paclitaxel] to [Paclitaxel + Trstuzumab] were censored at the time of crossover for the ***cumulative incidence*** analysis.

The cumulative proportions of [cardiac events] were estimated and compared by using the ***cumulative incidence*** function. The ***competing events*** were any first event of recurrence, second primary cancers, and deaths precluding [cardiac events]. The Cox ***cause-specific*** proportional hazards model was used to evaluate the association between time to [congestive heart failure] and cardiac risks factors. A parametric regression model on ***cause specific subdistribution*** hazard was used to build a prediction model for 5-year probability of developing [cardiac events] with 95% point-wise CIs, adjusting for significant risk factors.

These five patients were censored at the time of first dose of trastuzumab for the ***cumulative incidence*** analysis… Among patients assigned to receive [Paclitaxel + Trstuzumab], 947 had follow-up information for the ***cumulative incidence*** analysis.

On the basis of parametric regression on ***cause-specific subdistribution*** hazard, a Cardiac Risk Score can be calculated as follows:

The ***cumulative incidence*** of [cardiac events] in the control arm 7 years after day 1 of cycle 5 was 1.3% (95% CI, 0.5% to 2.1%), and the ***cumulative incidence*** among trastuzumab-treated evaluable patients was 4.0% (95% CI, 2.8% to 5.2%).

In [National Surgical Adjuvent Breast and Bowel Project protocol B-31], the 7-year difference in ***cumulative incidence*** of protocol-defined [cardiac events] between the experimental and control arms is 2.7% (4.0% v 1.3%, respectively). In the Herceptin Adjuvant trial, which required for random assignment an [left venticular ejection fraction] 55% after completion of all chemotherapy and radiation, the ***cumulative incidence*** of severe [congestive heart failure] with trastuzumab therapy was only 0.8% at a median follow-up of 3.6 years.

(Travis et al., 2005)

We estimated this future risk, taking into account age and calendar year of [Hodgkin's lymphoma] diagnosis, [Hodgkin's lymphoma] treatment information, population breast cancer incidence rates, and ***competing causes*** of death.

To compute cumulative absolute risks of breast cancer, we used modified standardized incidence ratios to relate cohort breast cancer risks to those in the general population, enabling application of population-based breast cancer rates, and we allowed for ***competing risks*** by using population-based mortality rates in female [Hodgkin's lymphoma] survivors.

Estimates of the ***cumulative incidence*** of breast cancer after treatment for [Hodgkin's lymphoma] at age 30 years or younger have been sparse, inconsistent, and series specific, ranging from 4.2% to 34% at 20 – 25 years of followup. Moreover, most estimates have not taken into account the influence of ***competing causes*** of mortality, which can be substantial among [Hodgkin's lymphoma] patients, or the effect of alkylating agent therapy, which can lower subsequent breast cancer risk.

We also took into account age and calendar year of [Hodgkin's lymphoma] diagnosis, age at counselling, baseline breast cancer incidence rates, and ***competing causes*** of mortality.

Finally, combining information on external relative risks with data on population breast cancer incidence rates from the Surveillance, Epidemiology, and End Results Program and with [Surveillance, Epidemiology, and End Results] Program data on ***competing causes*** of death in [Hodgkin's lymphoma] survivors, we estimated cumulative absolute risk of breast cancer, as described in detail below.

Our risk estimates derived from a large international population-based study; projections take into account age and calendar year at [Hodgkin's lymphoma] diagnosis, time since treatment, and ***competing causes*** of mortality.

Mortality rates for ***competing risk*** calculations, stratified by calendar year period and age range at [Hodgkin's lymphoma] diagnosis, were similarly derived from U.S. population-based data.

In most studies in which the absolute excess risk of breast cancer among women treated for [Hodgkin's lymphoma] at age 30 years or younger have been presented, numbers of case patients are also small (range = 14 – 19 case patients), resulting in highly variable estimates, and ***competing risks*** are not considered.

Two recent investigations of breast cancer after childhood or adolescent [Hodgkin's lymphoma], however, accounted for ***competing causes*** of death. Among girls treated with mantle radiotherapy for [Hodgkin's lymphoma] before age 17 years, the ***cumulative incidence*** of all invasive breast cancer (27 unilateral case patients plus 12 patients with contralateral tumors = 39) was 5.6% (95% CI = 2.8 to 8.3) and 16.9% (95% CI = 9.4 to 24.5%) at 20 and 30 years of follow-up, respectively.

These risks are calculated by using general population Surveillance, Epidemiology, and End Results Program rates for breast cancer and national rates for ***competing causes*** of mortality.

These values are not corrected for ***competing causes*** of mortality and are thus slightly larger than comparable estimates of absolute risk in Tables 2 and 3.

The mortality rates from ***competing risks*** are assumed to be known with negligible random error.

17 APPENDIX IX

A sensitivity analysis was conducted to test the assumptions made in modelling the risk of antenatal adverse events. The proportional hazards assumption, made by both the Cox and Royston-Parmar models, was tested by incorporating interactions between prognostic factors and log-time. The results are given in Table 17.1:

Table 17.1: Significance (p-values) of interactions of prognostic factors and ln(time) in fitted prognostic models

	Cox proportional hazards model	**Royston-Parmar flexible parametric model**
Maternal age	*0.254*	*0.294*
Gestational age	*0.414*	*0.777*
Medical history	*0.676*	*0.564*
Systolic blood pressure	*0.178*	*0.010*
Platelet count	*0.058*	*0.181*
Serum creatinine	*0.960*	*0.561*
Antihypertensive treatment	*0.828*	*0.877*
Magnesium sulphate treatment	*0.009*	*<0.001*

Significant interactions were detected between magnesium sulphate treatment and ln(time) in both models (p-value: 0.009 Cox, <0.001 Royston-Parmar). This treatment is given to minimise the risk of eclampsia and prevent preterm labour, its effects significantly decrease over time. The interaction between SBP and ln(time) was found to be significant in the Royston-Parmar model (p-value 0.010), whereas the interaction between platelet count and ln(time) was found to be significant in the Cox model (p-value: 0.058).

18 APPENDIX X

A sensitivity analysis was conducted to test the assumptions made during model development. The proportional hazards assumption, made by both the cause-specific and subdistribution approaches, was tested by incorporating interactions between the included prognostic factors and log time, using a linear (one degree of freedom) spline function. The results are given in Table 17.1.

Table 18.1: Significance (p-values) of interactions of prognostic factors and ln(time) in fitted prognostic models

	Cause-specific model: delivery	**Cause-specific model: antenatal adverse events**	**Subdistribution model: antenatal adverse events**
Maternal age (years)	0.005	0.298	0.426
Gestational age (weeks)	0.447	0.488	0.462
Medical history*	0.642	0.699	0.221
Systolic blood pressure	0.003	0.177	N/A
Platelet count	0.101	0.145	N/A
Alanine amino transaminase	0.511	N/A	N/A
Serum creatinine	0.652	0.838	N/A
Antihypertensive treatment	0.562	0.809	0.321
Magnesium sulphate treatment	<0.001	<0.001	<0.001

19 APPENDIX XI

The following describes the process for determining the constant cause-specific hazard functions for the simulations study conducted in Chapter 7:

The cause-specific cumulative incidence function $F_k(t)$ for event k at time t is written in terms of the overall survival function $\bar{F}(t)$ and the cause-specific hazard function $h_k(t)$:

$$F_k(t) = \int_0^t \bar{F}(s) h_k(s)\, ds \qquad \textit{Equation 19.1}$$

The overall survival function $\bar{F}(t)$ is written in terms of all cause-specific hazard functions, with two competing events $K = 2$:

$$\bar{F}(t) = \prod_{k=1}^{2} \exp\left\{-\int_0^t h_k(s)\, ds\right\} \qquad \textit{Equation 19.2}$$

Utilising exponential models with constant hazards $h_1(t) = \alpha_1$ for the event of interest and $h_2(t) = \alpha_2$, the overall survival function becomes:

$$\bar{F}(t) = \exp\{-\alpha_1 t\} \times \exp\{-\alpha_2 t\} = \exp\{-(\alpha_1 + \alpha_2)t\} \qquad \textit{Equation 19.3}$$

Incorporating this into the cause-specific cumulative incidence function for the event of interest $F_1(t)$, gives:

$$\begin{aligned} F_1(t) &= \int_0^t \alpha_1 \exp\{-(\alpha_1 + \alpha_2)s\}\, ds \\ &= \alpha_1 \left[\frac{\exp\{-(\alpha_1 + \alpha_2)t\}}{-(\alpha_1 + \alpha_2)}\right]_0^t \\ &= \alpha_1 \left[\frac{\exp\{-(\alpha_1 + \alpha_2)t\}}{-(\alpha_1 + \alpha_2)} - \frac{1}{-(\alpha_1 + \alpha_2)}\right] \end{aligned}$$

$$F_1(t) = \frac{\alpha_1}{\alpha_1 + \alpha_2}\left[1 - \exp\{-(\alpha_1 + \alpha_2)t\}\right] \qquad \textit{Equation 19.4}$$

Similarly, the cause-specific cumulative incidence for the competing event, $F_2(t)$ is:

$$F_2(t) = \frac{\alpha_2}{\alpha_1 + \alpha_2}\left[1 - \exp\{-(\alpha_1 + \alpha_2)t\}\right] \qquad \textit{Equation 19.5}$$

The relationship between the constant cause-specific hazards can be represented using the constant scaling factor γ:

$$\gamma \frac{\alpha_1}{\alpha_1+\alpha_2}[1 - exp\{-(\alpha_1 + \alpha_2)t\}] = \frac{\alpha_2}{\alpha_1+\alpha_2}[1 - exp\{-(\alpha_1 + \alpha_2)t\}] \therefore \gamma\alpha_1 = \alpha_2$$

Thus, given the proportion of participants who experience the event of interest $F_1(t)$, and the scaling factor γ, the constant cause-specific hazard term for both the event of interest and the competing event can be determined as follows:

$$F_1(t) = \int_0^t \alpha_1 \exp\{-(\alpha_1 + \alpha_2)s\}\, ds$$

$$= \alpha_1 \left[\frac{\exp\{-(\alpha_1+\alpha_2)t\}}{-(\alpha_1+\alpha_2)}\right]_0^t$$

$$= \alpha_1 \left[\frac{\exp\{-(\alpha_1+\alpha_2)t\}}{-(\alpha_1+\alpha_2)} - \frac{1}{-(\alpha_1+\alpha_2)}\right]$$

$$F_1(t) = \frac{\alpha_1}{\alpha_1+\alpha_2}[1 - \exp\{-(\alpha_1 + \alpha_2)t\}] \qquad \textbf{\textit{Equation 19.6}}$$

20 APPENDIX XII

Methods for generating survival times for time-to-event analyses are summarised by Bender et al. (Bender et al., 2005). Briefly, the hazard function of a proportional hazards model, $h(t|\mathbf{X})$, may be expressed as a function of the baseline hazard function, $h(t|\mathbf{0})$; a vector of prognostic factors, $\mathbf{X}$; and corresponding regression coefficients, $\boldsymbol{\beta}$; as:

$$h(t|\mathbf{X}) = h(t|\mathbf{0})\exp(\mathbf{X}\boldsymbol{\beta}) \qquad \textit{Equation 20.1}$$

The cumulative hazard function, $H(t|\mathbf{X})$, survival function, $S(t|\mathbf{X})$, and cumulative distribution function, $F(t|\mathbf{X})$, are expressed as:

$$H(t|\mathbf{X}) = H(t|\mathbf{0})\exp(\mathbf{X}\boldsymbol{\beta}), \text{ in which } H(t|\mathbf{0}) = \int_0^t h(u|\mathbf{0})du \qquad \textit{Equation 20.2}$$

$$S(t|\mathbf{X}) = \exp[-H(t|\mathbf{X})] \qquad \textit{Equation 20.3}$$

$$F(t|\mathbf{X}) = 1 - \exp[-H(t|\mathbf{X})] \qquad \textit{Equation 20.4}$$

Bender et al. (Bender et al., 2005) showed if T is the simulated survival time, then by letting:

$$F(T|\mathbf{X}) = 1 - \exp[-H(T|\mathbf{X})] = u, \text{ where } u \sim \text{Uniform}(0,1) \qquad \textit{Equation 20.5}$$

We can write;

$$S(T|\mathbf{X}) = \exp[-H(T|\mathbf{X})] = u \qquad \textit{Equation 20.6}$$

If $h(T|\mathbf{0}) > 0$ and $H(t|\mathbf{0})$ can be directly inverted, the above equation can be rearranged and solved fo T:

$$T = H^{-1}(-\ln(U)\exp(-\mathbf{X}\boldsymbol{\beta})|\mathbf{0}) \qquad \textit{Equation 20.7}$$

21 APPENDIX XIII

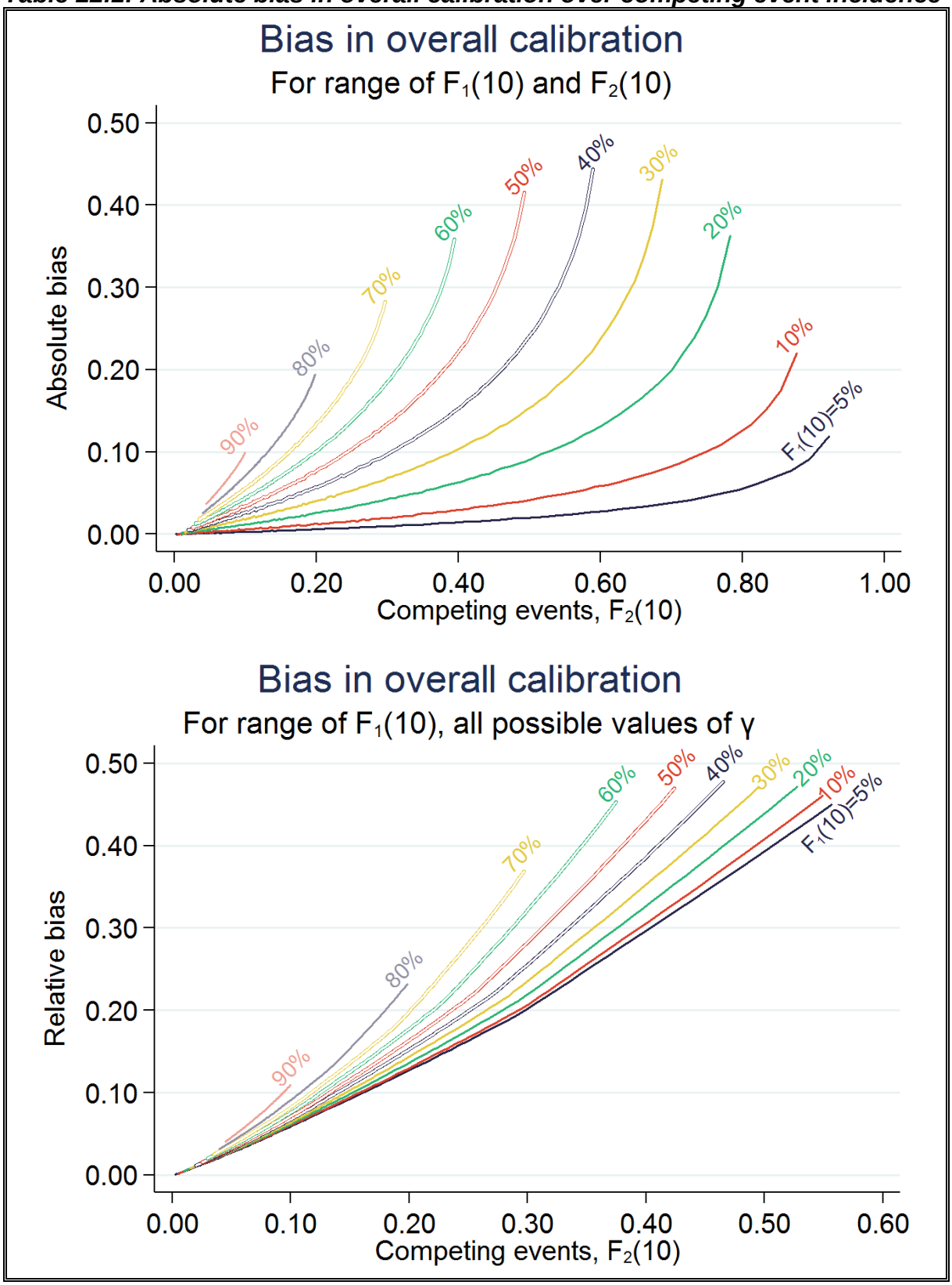
An example of Stata code used to generate data for the base scenario and evaluate overall calibration bias.

```
*****************************************************************
*     COMPETING RISKS PROGNOSTIC MODELS     *
*     SIMULATION STUDY   *
*****************************************************************

*(1)* Generate data and overall calibration bias for 500 studies
*Base scenarios 5-8: N=4144 participants, Exponential, EOI= 11.2%
*Gamma = 1.0, 1.5, 2.0, 2.72

*Set working directory
cd "C:\Users\rpd97\Desktop\Stata Simulations"

*Set seed for reproducibility
set seed 545245

*Set up loop to run through all base scenarios
local Scen = 4
foreach lam in "0.01449 0.03946" "0.01365 0.02730" ///
    "0.01314 0.01971" "0.01268 0.01268" {
      local Scen = `Scen' + 1

*Set expected cumulative incidence for time t=10*
      local CIF = 0.112

*Create a post file to save bias estimates to after each study
      postfile Scenario`Scen' Sim_No KM_Est Bias ///
using "Scenario`Scen'", replace

*Loop through 500 studies
      forvalues i = 1/500 {
            clear
*Create 4,144 participants
            set obs 4144

*Simulate the time and event data
            survsim Time Event, cr ncr(2) maxtime(10) ///
                  distribution(exponential) lambdas(`lam')

*Survival set the data
            stset Time, failure(Event==1)

*Obtain Kaplan-Meier estimates*
            sts gen surv = s
            sum surv if Event==1

*Calculate Kaplan-Meier and Bias estimates at time t=10
local KMest = 1 - r(min)
            local Bias = `KMest' - `CIF'

*Save bias estimates for each study to the post file
```

```
            post Scenario`Scen' (`i') (`KMest') (`Bias')
}
postclose Scenario`Scen'
}

*(2)* Histograms of distribution of overall calibration bias

*Set up loop to run through all base scenarios
forvalues i = 5/8 {
      use "Scenario`i'", clear

*generate histograms for each scenario
hist Bias, percent lcolor("39 36 73") fcolor("39 36 73") ///
fintensity(inten50) bin(10) graphregion(color(white))     /// title("Scenario
`j'") ylabel(0 (5) 25, angle(0))              ///
xscale(range(-0.03 0.15)) xlabel(-0.03 (0.03) 0.15)       /// yscale(range(0
25)) scale(1.2) saving("Graph`j'", replace)
}

*Combine all histograms into one plot
graph combine "Graph5" "Graph6" "Graph7" "Graph8",            ///
scale(0.8) graphregion(color(white)) rows(4)            ///
title("Distribution of overall calibration bias", span)

*(3)* Summary statistics of overall calibration bias

*Create a post file to save bias estimates to after each study
postfile Calibation_bias Scenario_No KM_mean KM_sd Abs_Bias_m ///
Abs_Bias_sd Rel_Bias_m Rel_Bias_sd                         ///
using "Calibration_bias", replace

*Set up loop to run through base scenarios
forvalues k = 5/8 {
      use " Scenario`k'", clear
*Summarise Kaplan-Meier failure estimates
      sum KM_Est
      local kmmean = r(mean)
      local kmsd = r(sd)

*Summarise absolute bias estimates
      sum Bias
      local absmean = r(mean)
      local abssd = r(sd)

*Generate and summarise relative bias estimates
      local CIF = 0.112
      gen Rel_Bias = Bias / `CIF'
      sum Rel_Bias
      local relmean = r(mean)
      local relsd = r(sd)

*Save bias summary statistics for each study to the post file
      post Calibation_bias (`k') (`kmmean') (`kmsd') (`absmean')///
(`abssd') (`relmean') (`relsd')
}
postclose Calibation_bias
```

22 APPENDIX XIV

Table 22.1: Absolute bias in overall calibration over competing event incidence



23 REFERENCES

ALLIGNOL, A., SCHUMACHER, M., WANNER, C., DRECHSLER, C. & BEYERSMANN, J. 2011. Understanding competing risks: a simulation point of view. *BMC medical research methodology,* 11**,** 1.

ALTMAN, D. G. 2001. Systematic reviews in health care: Systematic reviews of evaluations of prognostic variables. *BMJ: British Medical Journal,* 323**,** 224.

ALTMAN, D. G. & ROYSTON, P. 2000. What do we mean by validating a prognostic model? *Statistics in medicine,* 19**,** 453-473.

ALTMAN, D. G., VERGOUWE, Y., ROYSTON, P. & MOONS, K. G. 2009. Prognosis and prognostic research: validating a prognostic model. *Bmj,* 338**,** b605.

ANDERSEN, P. K., GESKUS, R. B., DE WITTE, T. & PUTTER, H. 2012. Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology,* 41**,** 861-870.

ANDRINOPOULOU, E.-R., RIZOPOULOS, D., TAKKENBERG, J. J. & LESAFFRE, E. 2017. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical methods in medical research,* 26**,** 1787-1801.

ARIESEN, M., ALGRA, A., WARLOW, C. & ROTHWELL, P. 2006. Predictors of risk of intracerebral haemorrhage in patients with a history of TIA or minor ischaemic stroke. *Journal of Neurology, Neurosurgery & Psychiatry,* 77**,** 92-94.

AUSTIN, P. C. & FINE, J. P. 2017. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in Medicine,* 36**,** 1203-1209.

AYERBE, L., GONZALEZ, E., GALLO, V., COLEMAN, C. L., WRAGG, A. & ROBSON, J. 2016. Clinical assessment of patients with chest pain; a systematic review of predictive tools. *BMC Cardiovasc Disord,* 16**,** 18.

BARTLETT, J. W. & TAYLOR, J. M. 2016. Missing covariates in competing risks analysis. *Biostatistics,* 17**,** 751-763.

BENDER, R., AUGUSTIN, T. & BLETTNER, M. 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine,* 24**,** 1713-1723.

BERNOULLI, D. 1760. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad. Roy. Sci.(Paris) avec Mém. des Math. et Phys. and Mém***,** 1-45.

BERRY, S. D., NGO, L., SAMELSON, E. J. & KIEL, D. P. 2010. Competing risk of death: an important consideration in studies of older adults. *Journal of the American Geriatrics Society,* 58**,** 783-787.

BEVILACQUA, J. L. B., KATTAN, M. W., CHANGHONG, Y., KOIFMAN, S., MATTOS, I. E., KOIFMAN, R. J. & BERGMANN, A. 2012. Nomograms for predicting the risk of arm lymphedema after axillary dissection in breast cancer. *Annals of surgical oncology,* 19**,** 2580-2589.

BEYERSMANN, J., LATOUCHE, A., BUCHHOLZ, A. & SCHUMACHER, M. 2009. Simulating competing risks data in survival analysis. *Statistics in medicine,* 28**,** 956-971.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., MOHER, D., RENNIE, D., DE VET, H. C. & LIJMER, J. G. 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of internal medicine,* 138**,** W1-12.

BRESLOW, N. E. 1972. Contribution to the discussion of the paper by DR Cox. *Journal of the Royal Statistical Society, Series B,* 34**,** 216-217.

BRIGANTI, A., GALLINA, A., SUARDI, N., CAPITANIO, U., TUTOLO, M., BIANCHI, M., PASSONI, N., SALONIA, A., COLOMBO, R. & DI GIROLAMO, V. 2010. Predicting erectile function recovery after bilateral nerve sparing radical prostatectomy: a proposal of a novel preoperative risk stratification. *The journal of sexual medicine,* 7**,** 2521-2531.

BRUNELLI, V. B. & PREFUMO, F. 2015. Quality of first trimester risk prediction models for pre-eclampsia: a systematic review. *Bjog,* 122**,** 904-14.

BURTON, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2006. The design of simulation studies in medical statistics. *Statistics in medicine,* 25**,** 4279-4292.

BUUREN, S. V. & OUDSHOORN, C. 2000. Multivariate imputation by chained equations: MICE V1. 0 user's manual. TNO.

CARAGATA, R., WYSSUSEK, K. H. & KRUGER, P. 2016. Acute kidney injury following liver transplantation: a systematic review of published predictive models. *Anaesth Intensive Care,* 44**,** 251-61.

CAREY, E. C., COVINSKY, K. E., LUI, L. Y., ENG, C., SANDS, L. P. & WALTER, L. C. 2008. Prediction of mortality in community-living frail elderly people with long-term care needs. *Journal of the American Geriatrics Society,* 56**,** 68-75.

CHARLSON, M. E., POMPEI, P., ALES, K. L. & MACKENZIE, C. R. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases,* 40**,** 373-383.

CHEONG-SEE, F., ALLOTEY, J., MARLIN, N., MOL, B., SCHUIT, E., RIET, G., RILEY, R., MOONS, K., KHAN, K. & THANGARATINAM, S. 2016. Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. *BJOG: An International Journal of Obstetrics & Gynaecology*.

COLLINS, G. S., MALLETT, S., OMAR, O. & YU, L.-M. 2011. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine,* 9**,** 103.

COLLINS, G. S., OGUNDIMU, E. O. & ALTMAN, D. G. 2016. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in medicine,* 35**,** 214-226.

COLLINS, G. S., OMAR, O., SHANYINDE, M. & YU, L.-M. 2013. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *Journal of clinical epidemiology,* 66**,** 268-277.

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. & MOONS, K. G. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC medicine,* 13**,** 1.

COUNSELL, C., DENNIS, M., MCDOWALL, M. & WARLOW, C. 2002. Predicting outcome after acute and subacute stroke development and validation of new prognostic models. *Stroke,* 33**,** 1041-1047.

COX, D. R. 1972a. Regression Models and Life-Tables *Journal of the Royal Statistical Society. Series B (Methodological),* 34**,** 187-220.

COX, D. R. 1972b. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society,* 34**,** 187-220.

CROWTHER, M. J. & LAMBERT, P. 2019. MULTISTATE: Stata module to perform multi-state survival analysis.

CROWTHER, M. J. & LAMBERT, P. C. 2013. Simulating biologically plausible complex survival data. *Statistics in medicine,* 32**,** 4118-4134.

CROWTHER, M. J. & LAMBERT, P. C. 2017. Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Statistics in medicine,* 36**,** 4719-4742.

CUSCHIERI, J. R., DRAWZ, P., FALCK-YTTER, Y. & WONG, R. C. 2014. Risk factors for acute gastrointestinal bleeding following myocardial infarction in veteran patients who are prescribed clopidogrel. *Journal of digestive diseases,* 15**,** 195-201.

DAMEN, J. A., HOOFT, L., SCHUIT, E., DEBRAY, T. P., COLLINS, G. S., TZOULAKI, I., LASSALE, C. M., SIONTIS, G. C., CHIOCCHIA, V., ROBERTS, C., SCHLUSSEL, M. M., GERRY, S., BLACK, J. A., HEUS, P., VAN DER SCHOUW, Y. T., PEELEN, L. M. & MOONS, K. G. 2016. Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj,* 353**,** i2416.

DE WREEDE, L. C., FIOCCO, M. & PUTTER, H. mstate: an R package for the analysis of competing risks and multi-state models.

DEBRAY, T. P., DAMEN, J. A., SNELL, K. I., ENSOR, J., HOOFT, L., REITSMA, J. B., RILEY, R. D. & MOONS, K. G. 2017. A guide to systematic review and meta-analysis of prediction model performance. *Bmj,* 356**,** i6460.

DEBRAY, T. P., VERGOUWE, Y., KOFFIJBERG, H., NIEBOER, D., STEYERBERG, E. W. & MOONS, K. G. 2015. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of clinical epidemiology,* 68**,** 279-289.

DIGNAM, J. J., ZHANG, Q. & KOCHERGINSKY, M. 2012. The use and interpretation of competing risks regression models. *Clinical Cancer Research,* 18**,** 2301-2308.

ECHOUFFO-TCHEUGUI, J. B., GREENE, S. J., PAPADIMITRIOU, L., ZANNAD, F., YANCY, C. W., GHEORGHIADE, M. & BUTLER, J. 2015. Population risk prediction models for incident heart failure: a systematic review. *Circ Heart Fail,* 8**,** 438-47.

EFRON, B. 1977. The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association,* 72**,** 557-565.

ENSOR, J., RILEY, R. D., JOWETT, S., MONAHAN, M., SNELL, K., BAYLISS, S., MOORE, D. & FITZMAURICE, D. 2016. Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation. *Health Technol Assess,* 20**,** i-xxxiii, 1-190.

EZAZ, G., LONG, J. B., GROSS, C. P. & CHEN, J. 2014. Risk prediction model for heart failure and cardiomyopathy after adjuvant trastuzumab therapy for breast cancer. *Journal of the American Heart Association,* 3**,** e000472.

FINE, J. P. & GRAY, R. J. 1999. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *American Statistical Association,* 94**,** 496-509.

FIOCCO, M., PUTTER, H. & VAN HOUWELINGEN, H. C. 2008. Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine,* 27**,** 4340-4358.

FOEKENS, J. A., PETERS, H. A., LOOK, M. P., PORTENGEN, H., SCHMITT, M., KRAMER, M. D., BRÜNNER, N., JÄNICKE, F., MEIJER-VAN GELDER, M. E. & HENZEN-LOGMANS, S. C. 2000. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research,* 60**,** 636-643.

FOX, K. A., DABBOUS, O. H., GOLDBERG, R. J., PIEPER, K. S., EAGLE, K. A., VAN DE WERF, F., AVEZUM, Á., GOODMAN, S. G., FLATHER, M. D. & ANDERSON, F. A. 2006. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). *bmj,* 333**,** 1091.

GA WELLS, B. S., D O'CONNELL, J PETERSON, V WELCH, M LOSOS, P TUGWELL,. 2014. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses* [Online]. Available: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [Accessed 27 June 2017].

GANZEVOORT, W., REP, A., BONSEL, G. J., FETTER, W. P., VAN SONDEREN, L., DE VRIES, J. I., WOLF, H. & INVESTIGATORS, P. 2005. A randomised controlled trial comparing two temporising management strategies, one with and one without plasma volume expansion, for severe and early onset pre-eclampsia. *BJOG,* 112**,** 1358-68.

GERDS, T. A., ANDERSEN, P. K. & KATTAN, M. W. 2014. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in medicine,* 33**,** 3191-3203.

GESKUS, R. B. 2015. *Data analysis with competing risks and intermediate states*, CRC Press.

GRAMBAUER, N., SCHUMACHER, M. & BEYERSMANN, J. 2010. Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in medicine,* 29**,** 875-884.

GRAMBSCH, P. M. & THERNEAU, T. M. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika,* 81**,** 515-526.

GRAY, E. P., TEARE, M. D., STEVENS, J. & ARCHER, R. 2016. Risk Prediction Models for Lung Cancer: A Systematic Review. *Clin Lung Cancer,* 17**,** 95-106.

GROENWOLD, R. H., MOONS, K. G., PAJOUHESHNIA, R., ALTMAN, D. G., COLLINS, G. S., DEBRAY, T. P., REITSMA, J. B., RILEY, R. D. & PEELEN, L. M. 2016. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of clinical epidemiology,* 78**,** 90-100.

HARRELL, F. E. 2015. *Regression Modelling Strategies*, Springer International Publishing.

HARRELL, F. E., CALIFF, R. M., PRYOR, D. B., LEE, K. L. & ROSATI, R. A. 1982. Evaluating the yield of medical tests. *Jama,* 247**,** 2543-2546.

HASKINS, R., OSMOTHERLY, P. G. & RIVETT, D. A. 2015. Validation and impact analysis of prognostic clinical prediction rules for low back pain is needed: a systematic review. *J Clin Epidemiol,* 68**,** 821-32.

HAYBITTLE, J., BLAMEY, R., ELSTON, C., JOHNSON, J., DOYLE, P., CAMPBELL, F., NICHOLSON, R. & GRIFFITHS, K. 1982. A prognostic index in primary breast cancer. *British journal of cancer,* 45**,** 361.

HAYDEN, J. A., VAN DER WINDT, D. A., CARTWRIGHT, J. L., CÔTÉ, P. & BOMBARDIER, C. 2013. Assessing bias in studies of prognostic factors. *Annals of internal medicine,* 158**,** 280-286.

HEINZE, G., WALLISCH, C. & DUNKLER, D. 2018. Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal,* 60**,** 431-449.

HEMINGWAY, H., CROFT, P., PEREL, P., HAYDEN, J. A., ABRAMS, K., TIMMIS, A., BRIGGS, A., UDUMYAN, R., MOONS, K. G. & STEYERBERG, E. W. 2013. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes.

HEMINGWAY, H., RILEY, R. D. & ALTMAN, D. G. 2009. Ten steps towards improving prognosis research. *Bmj,* 339**,** b4184.

HILKENS, N. A., ALGRA, A. & GREVING, J. P. 2016. Prediction models for intracranial hemorrhage or major bleeding in patients on antiplatelet therapy: a systematic review and external validation study. *J Thromb Haemost,* 14**,** 167-74.

HINCHLIFFE, S. R. 2013. *Advancing and appraising competing risks methodology for better communication of survival statistics.* University of Leicester.

HINCHLIFFE, S. R. & LAMBERT, P. C. 2013a. Extending the flexible parametric survival model for competing risks. *The Stata Journal,* 13**,** 344-355.

HINCHLIFFE, S. R. & LAMBERT, P. C. 2013b. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC medical research methodology,* 13**,** 1.

HINCHLIFFE, S. R., SEATON, S. E., LAMBERT, P. C., DRAPER, E. S., FIELD, D. J. & MANKTELOW, B. N. 2013. Modelling time to death or discharge in neonatal care: an application of competing risks. *Paediatric and perinatal epidemiology,* 27**,** 426-433.

HINGORANI, A. D., VAN DER WINDT, D. A., RILEY, R. D., ABRAMS, K., MOONS, K. G., STEYERBERG, E. W., SCHROTER, S., SAUERBREI, W., ALTMAN, D. G. & HEMINGWAY, H. 2013. Prognosis research strategy (PROGRESS) 4: stratified medicine research.

HIPPISLEY-COX, J. & COUPLAND, C. 2012. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract,* 62**,** e29-e37.

INGUI, B. J. & ROGERS, M. A. 2001. Searching for clinical prediction rules in MEDLINE. *Journal of the American Medical Informatics Association,* 8**,** 391-397.

JANSSEN, K., MOONS, K., KALKMAN, C., GROBBEE, D. & VERGOUWE, Y. 2008. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of clinical epidemiology,* 61**,** 76-86.

JAVANMARDI, F., SAKI-MALEHI, A., AHMADZADEH, A. & RAHIM, F. 2018. Assessing Prognostic Factors in Hodgkin's Lymphoma: Multistate Illness-Death Model. *International journal of hematology-oncology and stem cell research,* 12**,** 57.

JENKINS, D. A., SPERRIN, M., MARTIN, G. P. & PEEK, N. 2018. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnostic and prognostic research,* 2**,** 23.

KAPLAN, E. L. & MEIER, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association,* 53**,** 457-481.

KOHN, C. G., MEARNS, E. S., PARKER, M. W., HERNANDEZ, A. V. & COLEMAN, C. I. 2015. Prognostic accuracy of clinical prediction rules for early post-pulmonary embolism all-cause mortality: a bivariate meta-analysis. *Chest,* 147**,** 1043-62.

KOLLER, M. T., RAATZ, H., STEYERBERG, E. W. & WOLBERS, M. 2012. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in medicine,* 31**,** 1089-1097.

KOVALCHIK, S. A., RONCKERS, C. M., VEIGA, L. H., SIGURDSON, A. J., INSKIP, P. D., DE VATHAIRE, F., SKLAR, C. A., DONALDSON, S. S., ANDERSON, H. & BHATTI, P. 2012. Absolute risk prediction of second primary thyroid cancer among 5-year survivors of childhood cancer. *Journal of Clinical Oncology,* 31**,** 119-127.

LAMBERT, P. C. *Competing Risks* [Online]. Available: https://pclambert.net/interactivegraphs/competing_risks/competingrisks [Accessed 16/06/2019 2019].

LAMBERT, P. C. & ROYSTON, P. 2009. Further development of flexible parametric models for survival analysis. *The Stata Journal,* 9**,** 265-290.

LAMBERT, P. C., WILKES, S. R. & CROWTHER, M. J. 2017. Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in medicine,* 36**,** 1429-1446.

LAMBIE, M., TEECE, L., JOHNSON, D., PETRIE, M., MACTIER, R., SOLIS-TRAPALA, I., BELCHER, J., BEKKER, H., WILKIE, M. & TUPLING, K. 2019. Estimating risk of encapsulating peritoneal sclerosis accounting for the competing risk of death. *Nephrology Dialysis Transplantation*.

LATOUCHE, A., BOISSON, V., CHEVRET, S. & PORCHER, R. 2007. Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in medicine,* 26**,** 965-974.

LAU, B., COLE, S. R. & GANGE, S. J. 2009. Competing risk regression models for epidemiologic data. *American journal of epidemiology***,** kwp107.

LEOCE, N. M. 2016. *Prognostic Modeling in the Presence of Competing Risks: An Application to Cardiovascular and Cancer Mortality in Breast Cancer Survivors*, Columbia University.

LI, B., CAIRNS, J. A., ROBB, M. L., JOHNSON, R. J., WATSON, C. J., FORSYTHE, J. L., ONISCU, G. C., RAVANAN, R., DUDLEY, C. & RODERICK, P. 2016. Predicting patient survival after deceased donor kidney transplantation using flexible parametric modelling. *BMC nephrology,* 17**,** 51.

LIM, C., DEJONG, C. H. & FARGES, O. 2015. Improving the quality of liver resection: a systematic review and critical analysis of the available prognostic models. *HPB (Oxford),* 17**,** 209-21.

LINSELL, L., MALOUF, R., JOHNSON, S., MORRIS, J., KURINCZUK, J. J. & MARLOW, N. 2016. Prognostic Factors for Behavioral Problems and Psychiatric Disorders in Children Born Very Preterm or Very Low Birth Weight: A Systematic Review. *J Dev Behav Pediatr,* 37**,** 88-102.

LITTLE, R. J. 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics,* 6**,** 287-296.

LUNN, M. & MCNEIL, D. 1995. Applying Cox regression to competing risks. *Biometrics***,** 524-532.

LUO, G., NKOY, F. L., STONE, B. L., SCHMICK, D. & JOHNSON, M. D. 2015. A systematic review of predictive models for asthma development in children. *BMC Med Inform Decis Mak,* 15**,** 99.

MAHAJERIN, A., BRANCHFORD, B. R., AMANKWAH, E. K., RAFFINI, L., CHALMERS, E., VAN OMMEN, C. H. & GOLDENBERG, N. A. 2015. Hospital-associated venous thromboembolism in pediatrics: a systematic review and meta-analysis of risk factors and risk-assessment models. *Haematologica,* 100**,** 1045-50.

MAO, Z., ZHONG, X., YIN, J., ZHAO, Z., HU, X. & HACKETT, M. L. 2015. Predictors associated with stroke after coronary artery bypass grafting: a systematic review. *J Neurol Sci,* 357**,** 1-7.

MARQUES, A., FERREIRA, R. J., SANTOS, E., LOZA, E. & CARMONA, L. 2015. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. 74**,** 1958-67.

MARSHALL, A., ALTMAN, D. G., HOLDER, R. L. & ROYSTON, P. 2009. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC medical research methodology,* 9**,** 1.

MARUFU, T. C., MANNINGS, A. & MOPPETT, I. K. 2015. Risk scoring models for predicting peri-operative morbidity and mortality in people with fragility hip fractures: Qualitative systematic review. *Injury,* 46**,** 2325-34.

MATHIEU, R., ARANGO, J. D. O., BECKENDORF, V., DELOBEL, J.-B., MESSAI, T., CHIRA, C., BOSSI, A., LE PRISÉ, E., GUERIF, S. & SIMON, J.-M. 2014. Nomograms to predict late urinary toxicity after prostate cancer radiotherapy. *World journal of urology,* 32**,** 743-751.

MAZZAFERRO, V., SPOSITO, C., ZHOU, J., PINNA, A. D., DE CARLIS, L., FAN, J., CESCON, M., DI SANDRO, S., YI-FENG, H. & LAUTERIO, A. 2018. Metroticket 2.0 model for analysis of competing risks of death after liver transplantation for hepatocellular carcinoma. *Gastroenterology,* 154**,** 128-139.

MCGINN, T. G., GUYATT, G. H., WYER, P. C., NAYLOR, C. D., STIELL, I. G., RICHARDSON, W. S. & GROUP, E.-B. M. W. 2000. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. *Jama,* 284**,** 79-84.

MEYER, M. J., PEREIRA, S., MCCLURE, A., TEASELL, R., THIND, A., KOVAL, J., RICHARDSON, M. & SPEECHLEY, M. 2015. A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation. *Disabil Rehabil,* 37**,** 1316-23.

MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. G. & GROUP, P. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med,* 6**,** e1000097.

MOONS, K. G., ALTMAN, D. G., REITSMA, J. B., IOANNIDIS, J. P., MACASKILL, P., STEYERBERG, E. W., VICKERS, A. J., RANSOHOFF, D. F. & COLLINS, G. S. 2015. Transparent Reporting of a multivariable prediction model for individual Prognosis or diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: explanation and elaboration. *Annals of internal medicine,* 162**,** W1-W73.

MOONS, K. G., ALTMAN, D. G., VERGOUWE, Y. & ROYSTON, P. 2009a. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj,* 338**,** b606.

MOONS, K. G., COOK, N. & COLLINS, G. 2017. A new community for those involved and interested in diagnosis and prognosis. BioMed Central.

MOONS, K. G., DE GROOT, J. A., BOUWMEESTER, W., VERGOUWE, Y., MALLETT, S., ALTMAN, D. G., REITSMA, J. B. & COLLINS, G. S. 2014. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS medicine,* 11**,** e1001744.

MOONS, K. G., KENGNE, A. P., GROBBEE, D. E., ROYSTON, P., VERGOUWE, Y., ALTMAN, D. G. & WOODWARD, M. 2012a. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart,* 98**,** 691-698.

MOONS, K. G., KENGNE, A. P., WOODWARD, M., ROYSTON, P., VERGOUWE, Y., ALTMAN, D. G. & GROBBEE, D. E. 2012b. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart,* 98**,** 683-690.

MOONS, K. G., ROYSTON, P., VERGOUWE, Y., GROBBEE, D. E. & ALTMAN, D. G. 2009b. Prognosis and prognostic research: what, why, and how? *Bmj,* 338**,** b375.

MORRIS, T. P., WHITE, I. R., CARPENTER, J. R., STANWORTH, S. J. & ROYSTON, P. 2015. Combining fractional polynomial model building with multiple imputation. *Statistics in medicine,* 34**,** 3298-3317.

MORRIS, T. P., WHITE, I. R. & CROWTHER, M. J. 2019. Using simulation studies to evaluate statistical methods. *Statistics in medicine*.

MOZUMDER, S. I. *InterPreT Cancer Survival* [Online]. Available: https://interpret.le.ac.uk [Accessed 10/06/2019 2019].

NAKAGAWA, Y., SANNOMIYA, K., KINOSHITA, M., SHIOMI, T., OKADA, K., YOKOYAMA, H., SAWAGUTI, Y., MINAMOTO, K., WEI, C.-N. & OHMORI, S. 2008. Development of an assessment sheet for fall prediction in stroke inpatients in convalescent rehabilitation wards in Japan. *Environmental health and preventive medicine,* 13**,** 138-147.

NICOLAIE, M., VAN HOUWELINGEN, J., DE WITTE, T. & PUTTER, H. 2013. Dynamic prediction by landmarking in competing risks. *Statistics in medicine,* 32**,** 2031-2047.

NOORDZIJ, M., LEFFONDRÉ, K., VAN STRALEN, K. J., ZOCCALI, C., DEKKER, F. W. & JAGER, K. J. 2013. When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation,* 28**,** 2670-2677.

NYBERG, L. & GUSTAFSON, Y. 1997. Fall prediction index for patients in stroke rehabilitation. *Stroke,* 28**,** 716-721.

O'CAOIMH, R., CORNALLY, N., WEATHERS, E., O'SULLIVAN, R., FITZGERALD, C., ORFILA, F., CLARNETTE, R., PAUL, C. & MOLLOY, D. W. 2015. Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults. *Maturitas,* 82**,** 3-21.

OLIVER, C. M., WALKER, E., GIANNARIS, S., GROCOTT, M. P. & MOONESINGHE, S. R. 2015. Risk assessment tools validated for patients undergoing emergency laparotomy: a systematic review. *Br J Anaesth,* 115**,** 849-60.

ORGANIZATION, W. H. 1994. *Health care for the elderly: a manual for primary health care workers*.

PAGE, M. J., SHAMSEER, L., ALTMAN, D. G., TETZLAFF, J., SAMPSON, M., TRICCO, A. C., CATALÁ-LÓPEZ, F., LI, L., REID, E. K. & SARKIS-ONOFRE, R. 2016. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. *PLoS medicine,* 13**,** e1002028.

PEDUZZI, P., CONCATO, J., FEINSTEIN, A. R. & HOLFORD, T. R. 1995. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of clinical epidemiology,* 48**,** 1503-1510.

PEDUZZI, P., CONCATO, J., KEMPER, E., HOLFORD, T. R. & FEINSTEIN, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology,* 49**,** 1373-1379.

PINTILIE, M. 2007. Analysing and interpreting competing risk data. *Statistics in medicine,* 26**,** 1360-1367.

PROUST-LIMA, C. & TAYLOR, J. M. 2009. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics,* 10**,** 535-549.

PUTTER, H., FIOCCO, M. & GESKUS, R. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine,* 26**,** 2389-2430.

QUINLIVAN, L., COOPER, J., DAVIES, L., HAWTON, K., GUNNELL, D. & KAPUR, N. 2016. Which are the most useful scales for predicting repeat self-harm? A systematic review evaluating risk scales using measures of diagnostic accuracy. *BMJ Open,* 6**,** e009297.

REID, N. 1994. A conversation with sir david cox. *Statistical Science,* 9**,** 439-455.

RILEY, R., VAN DER WINDT, D., CROFT, P. & MOONS, K. G. 2019a. *Prognosis Research in Health Care: Concepts, Methods, and Impact*, Oxford University Press.

RILEY, R. D., ENSOR, J., SNELL, K. I., DEBRAY, T. P., ALTMAN, D. G., MOONS, K. G. & COLLINS, G. S. 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *Bmj,* 353**,** i3140.

RILEY, R. D., HAYDEN, J. A., STEYERBERG, E. W., MOONS, K. G., ABRAMS, K., KYZAS, P. A., MALATS, N., BRIGGS, A., SCHROTER, S. & ALTMAN, D. G. 2013. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med,* 10**,** e1001380.

RILEY, R. D., SNELL, K. I., ENSOR, J., BURKE, D. L., HARRELL JR, F. E., MOONS, K. G. & COLLINS, G. S. 2019b. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. *Statistics in Medicine,* 38**,** 1276-1296.

ROMOND, E. H., JEONG, J.-H., RASTOGI, P., SWAIN, S. M., GEYER JR, C. E., EWER, M. S., RATHI, V., FEHRENBACHER, L., BRUFSKY, A. & AZAR, C. A. 2012. Seven-year follow-up assessment of cardiac function in NSABP B-31, a randomized trial comparing doxorubicin and cyclophosphamide followed by paclitaxel (ACP) with ACP plus trastuzumab as adjuvant therapy for patients with node-positive, human epidermal growth factor receptor 2–positive breast cancer. *Journal of Clinical Oncology,* 30**,** 3792-3799.

ROYSTON, P. & ALTMAN, D. G. 2013. External validation of a Cox prognostic model: principles and methods. *BMC medical research methodology,* 13**,** 1.

ROYSTON, P. & LAMBERT, P. C. 2011. Flexible parametric survival analysis using Stata: beyond the Cox model.

ROYSTON, P., MOONS, K. G., ALTMAN, D. G. & VERGOUWE, Y. 2009. Prognosis and prognostic research: developing a prognostic model. *Bmj,* 338**,** b604.

ROYSTON, P. & PARMAR, M. K. 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine,* 21**,** 2175-2197.

ROYSTON, P. & SAUERBREI, W. 2004. A new measure of prognostic separation in survival data. *Statistics in medicine,* 23**,** 723-748.

ROYSTON, P. P., MAHESH K B; ALTMAN, DOUGLAS G 2010. External validation and updating of a prognostic survival model. https://www.ucl.ac.uk/drupal/site_statistics/sites/statistics/files/migrated-files/rr307.pdf: Department of Statistical Science at University College London.

RUBIN, D. B. 2004. *Multiple imputation for nonresponse in surveys*, John Wiley & Sons.

SALZ, T., BAXI, S. S., RAGHUNATHAN, N., ONSTAD, E. E., FREEDMAN, A. N., MOSKOWITZ, C. S., DALTON, S. O., GOODMAN, K. A., JOHANSEN, C., MATASAR, M. J., DE NULLY BROWN, P., OEFFINGER, K. C. & VICKERS, A. J. 2015. Are we ready to predict late effects? A systematic review of clinically useful prediction models. *Eur J Cancer,* 51**,** 758-66.

SCHATZKIN, A. & SLUD, E. 1989. Competing risks bias arising from an omitted risk factor. *American journal of epidemiology,* 129**,** 850-856.

SCHEMPER, M. & SMITH, T. L. 1996. A note on quantifying follow-up in studies of failure time. *Controlled clinical trials,* 17**,** 343-346.

SCHOENFELD, D. 1982. Partial residuals for the proportional hazards regression model. *Biometrika,* 69**,** 239-241.

SCHONBERG, M. A., DAVIS, R. B., MCCARTHY, E. P. & MARCANTONIO, E. R. 2009. Index to predict 5-year mortality of community-dwelling adults aged 65 and older using data from the National Health Interview Survey. *Journal of general internal medicine,* 24**,** 1115.

SCHUIT, E., GROENWOLD, R. H., HARRELL, F. E., DE KORT, W. L., KWEE, A., MOL, B. W. J., RILEY, R. D. & MOONS, K. G. 2013. Unexpected predictor–outcome associations in clinical prediction research: causes and solutions. *Canadian Medical Association Journal,* 185**,** E499-E505.

SCHUMACHER, M., OHNEBERG, K. & BEYERSMANN, J. 2016. Competing risk bias was common in a prominent medical journal. *Journal of Clinical Epidemiology,* 80**,** 135-136.

SCRUTINIO, D., GUIDA, P., PASSANTINO, A., AMMIRATI, E., OLIVA, F., LAGIOIA, R. & FRIGERIO, M. 2018. Application of competing risks analysis improved prognostic assessment of patients with decompensated chronic heart failure and reduced left ventricular ejection fraction. *Journal of clinical epidemiology,* 103**,** 31-39.

SEKHRI, N., FEDER, G. S., JUNGHANS, C., ELDRIDGE, S., UMAIPALAN, A., MADHU, R., HEMINGWAY, H. & TIMMIS, A. D. 2008. Incremental prognostic

value of the exercise electrocardiogram in the initial assessment of patients with suspected angina: cohort study. *Bmj,* 337**,** a2240.

SHEN, W., SAKAMOTO, N. & YANG, L. 2015. Cancer-specific mortality and competing mortality in patients with head and neck squamous cell carcinoma: a competing risk analysis. *Annals of surgical oncology,* 22**,** 264-271.

SIBAI, B. 2003. Diagnosis and Management of Gestational Hypertension and Preeclampsia. *Obstetrics & Gynecology,* 102**,** 181-192.

SILVER, S. A., SHAH, P. M., CHERTOW, G. M., HAREL, S., WALD, R. & HAREL, Z. 2015. Risk prediction models for contrast induced nephropathy: systematic review. *Bmj,* 351**,** h4395.

SILVERBERG, N. D., GARDNER, A. J., BRUBACHER, J. R., PANENKA, W. J., LI, J. J. & IVERSON, G. L. 2015. Systematic review of multivariable prognostic models for mild traumatic brain injury. *J Neurotrauma,* 32**,** 517-26.

SMIT, H. A., PINART, M., ANTO, J. M., KEIL, T., BOUSQUET, J., CARLSEN, K. H., MOONS, K. G., HOOFT, L. & CARLSEN, K. C. 2015. Childhood asthma prediction models: a systematic review. *Lancet Respir Med,* 3**,** 973-84.

SNELL, K. I. E. 2015. *Development and application of statistical methods for prognosis research.* University of Birmingham.

STEYERBERG, E. 2008. *Clinical prediction models: a practical approach to development, validation, and updating*, Springer Science & Business Media.

STEYERBERG, E. W., MOONS, K. G., VAN DER WINDT, D. A., HAYDEN, J. A., PEREL, P., SCHROTER, S., RILEY, R. D., HEMINGWAY, H., ALTMAN, D. G. & GROUP, P. 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med,* 10**,** e1001381.

TANG, E. Y., HARRISON, S. L., ERRINGTON, L., GORDON, M. F., VISSER, P. J., NOVAK, G., DUFOUIL, C., BRAYNE, C., ROBINSON, L., LAUNER, L. J. & STEPHAN, B. C. 2015. Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review. *PLoS One,* 10**,** e0136181.

THANGARATINAM, S., ALLOTEY, J., MARLIN, N., DODDS, J., CHEONG-SEE, F., VON DADELSZEN, P., GANZEVOORT, W., AKKERMANS, J., KERRY, S. & MOL, B. W. 2017. Prediction of complications in early-onset pre-eclampsia (PREP): development and external multinational validation of prognostic models. *BMC medicine,* 15**,** 68.

TRAVIS, L. B., HILL, D., DORES, G. M., GOSPODAROWICZ, M., VAN LEEUWEN, F. E., HOLOWATY, E., GLIMELIUS, B., ANDERSSON, M., PUKKALA, E. & LYNCH, C. F. 2005. Cumulative absolute breast cancer risk for young women treated for Hodgkin lymphoma. *Journal of the National Cancer Institute,* 97**,** 1428-1437.

USHER-SMITH, J. A., WALTER, F. M., EMERY, J. D., WIN, A. K. & GRIFFIN, S. J. 2016. Risk Prediction Models for Colorectal Cancer: A Systematic Review. *Cancer Prev Res (Phila),* 9**,** 13-26.

VAN WALRAVEN, C. & HAWKEN, S. 2016. Competing risk bias in Kaplan–Meier risk estimates can be corrected. *Journal of clinical epidemiology,* 70**,** 101-105.

VARADHAN, R., WEISS, C. O., SEGAL, J. B., WU, A. W., SCHARFSTEIN, D. & BOYD, C. 2010. Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications. *Medical care***,** S96-S105.

VERGOUWE, Y., STEYERBERG, E. W., EIJKEMANS, M. J. & HABBEMA, J. D. F. 2005. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of clinical epidemiology,* 58**,** 475-483.

VICKERS, A. J., VAN CALSTER, B. & STEYERBERG, E. W. 2015. Net benefit approaches to the evaluation of prediction models, molecular markers and diagnostic tests. *BMJ-British Medical Journal*.

VON DADELSZEN, P., MENZIES, J. M., PAYNE, B., MAGEE, L. A. & GROUP, P. S. 2009. Predicting adverse outcomes in women with severe pre-eclampsia. *Semin Perinatol,* 33**,** 152-7.

WALRAVEN, C. V. & MCALISTER, F. A. 2016. Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals. *Journal of clinical epidemiology,* 69**,** 170-173. e8.

WALSH, M. E., HORGAN, N. F., WALSH, C. D. & GALVIN, R. 2016. Systematic review of risk prediction models for falls after stroke. *J Epidemiol Community Health,* 70**,** 513-9.

WARNELL, I., CHINCHOLKAR, M. & ECCLES, M. 2015. Predicting perioperative mortality after oesophagectomy: a systematic review of performance and methods of multivariate models. *Br J Anaesth,* 114**,** 32-43.

WESSLER, B. S., PAULUS, J., LUNDQUIST, C. M., AJLAN, M., NATTO, Z., JANES, W. A., JETHMALANI, N., RAMAN, G., LUTZ, J. S. & KENT, D. M. 2017. Tufts PACE clinical predictive model registry: update 1990 through 2015. *Diagnostic and prognostic research,* 1**,** 20.

WHITE, I. R., DANIEL, R. & ROYSTON, P. 2010. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational statistics & data analysis,* 54**,** 2267-2275.

WHITE, I. R. & ROYSTON, P. 2009. Imputing missing covariate values for the Cox model. *Stat Med,* 28**,** 1982-1998.

WHITING, P., RUTJES, A. W., REITSMA, J. B., BOSSUYT, P. M. & KLEIJNEN, J. 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC medical research methodology,* 3**,** 25.

WHITING, P. F., RUTJES, A. W., WESTWOOD, M. E., MALLETT, S., DEEKS, J. J., REITSMA, J. B., LEEFLANG, M. M., STERNE, J. A. & BOSSUYT, P. M. 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine,* 155**,** 529-536.

WILKINSON, H. 2011. Saving mothers' lives. Reviewing maternal deaths to make motherhood safer: 2006–2008. *BJOG: An International Journal of Obstetrics & Gynaecology,* 118**,** 1402-1403.

WILLIAMS, T. G., CUBIELLA, J., GRIFFIN, S. J., WALTER, F. M. & USHER-SMITH, J. A. 2016. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. *BMC Gastroenterol,* 16**,** 63.

WILSON, T., QUAN, S., CHEEMA, K., ZARNKE, K., QUINN, R., DE KONING, L., DIXON, E., PANNU, N. & JAMES, M. T. 2016. Risk prediction models for acute kidney injury following major noncardiac surgery: systematic review. *Nephrol Dial Transplant,* 31**,** 231-40.

WOLBERS, M., BLANCHE, P., KOLLER, M. T., WITTEMAN, J. C. & GERDS, T. A. 2014. Concordance for prognostic models with competing risks. *Biostatistics,* 15**,** 526-539.

WOLBERS, M., KOLLER, M. T., WITTEMAN, J. C. & STEYERBERG, E. W. 2009. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology,* 20**,** 555-561.

WOLFF, R. F., MOONS, K. G., RILEY, R. D., WHITING, P. F., WESTWOOD, M., COLLINS, G. S., REITSMA, J. B., KLEIJNEN, J. & MALLETT, S. 2019. Probast: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine,* 170**,** 51-58.

WOLKEWITZ, M., COOPER, B. S., BONTEN, M. J., BARNETT, A. G. & SCHUMACHER, M. 2014. Interpreting and comparing risks in the presence of competing events. *BMJ,* 349**,** g5060.

WOOD, A. M., WHITE, I. R. & ROYSTON, P. 2008. How should variable selection be performed with multiply imputed data? *Statistics in medicine,* 27**,** 3227-3246.

WYATT, J. C. & ALTMAN, D. G. 1995. Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ,* 311**,** 1539-1541.